SAPIENZA UNIVERSITY OF ROME

DOCTORAL THESIS

Modelling Ovarian Follicle Dynamics within Assisted Reproductive Technology Treatments

Author:
Mariya MARKELOVA

Supervisor:
Enrico TRONCI

A thesis submitted in fulfillment of the requirements for the degree of Doctor of Philosophy

in the

Computer Science

January 7, 2018

Declaration of Authorship

I confirm that this PhD thesis is my own work and I have documented all sources and material used. This thesis was not previously presented to another examination board and has not been published.

Signed: 

Date: 07/01/2018

SAPIENZA UNIVERSITY OF ROME

Abstract

Faculty of Information Engineering, Informatics, and Statistics
Department of Computer Science

Doctor of Philosophy

Modelling Ovarian Follicle Dynamics within Assisted Reproductive Technology Treatments

by Mariya MARKELOVA

Infertility affects from 12% to 15% of reproductive couples in Western Europe. Most of infertility cases are related to female endocrinological problems and costs around 1 billion Euro per year. Assisted Reproduction Techniques have made huge improvements on chances of infertile couples. However, the success rate is drastically low.

Systems biology is an complex approach to tackle an entire organism, instead of singling out it's fractions and trying to understand them. The intention of this thesis is to apply systems biology to the problem of infertility.

Sufficient amount of research has been done towards design a whole-body model. However, none of them closely deal with endocrinological problems thus, they do not fully covers the problem of infertility. A great deal of work was done specifically oriented on recreating the dynamics of reproductive hormones. Such models have a high complexity and more than 100 parameters to be identified. Despite the ability to simulate concentration of hormones, the problem of identifing values for such a large amount of unknown parameters remains unresolved or highly complex.

Whereas models as (Röblitz et al., 2013) oriented on simulating the dynamics of multiple hormones such as Progesterone, Follicle-Stimulating Hormone, Luteinizing Hormone within normal cycle, this thesis oriented on establishing several models designed specifically for Estradiol concentration and follicle dynamics within stimulation treatment. Main aim is to reduce or eliminate number of measurements taken from a patient in order to increase patient comfort and reduce cost of a treatment.

This thesis was done within European Project PAEON, as a part of collaboration between Model Checking Group Laboratory (Sapienza University of Rome) and experts in reproductive medicine (Prof. Dr. med. Brigitte Leeners, University Hospital of Zürich).

Acknowledgements

I would like to offer my sincere gratitude to my advisor Prof. Enrico Tronci and to my family for their constant encouragements and support.

Contents

List of Figures

List of Tables

List of Abbreviations

TV-US	Trans-Vaginal Ultrasound
E2	Estradiol hormone
MR	External Factor (AFC) - Medium Response, see Section 2.2.4
ER	External Factor (AFC) - Elevated Response
HR	External Factor (AFC) - High Response
F/L	External Factor (Administered drug) - Follicle-Stimulating Hormone/Luteinizing Hormone
F	External Factor (Administered drug) - Follicle-Stimulating Hormone, see Section 2.2.4
I	External Factor (Health Condition) - Idiopathic, see Section 2.2.4
E	External Factor (Health Condition) - Endometriosis
O	External Factor (Health Condition) - Other
FP	Follicle Profile, see Section 2.1
ART	Assisted Reproduction Techniques
MC	Medical Case, see Section 2.1

Chapter 1

Introduction

1.1 Framework

According to (Wang et al., 2003), (Gnoth et al., 2003) infertility affects from 12% to 15% of reproductive couples and will drastically increase. It has a deep impact on our society, as (Daar and Merali, 2002) stated "the consequences of infertility in developing countries range from severe economic deprivation, to social isolation." Most of infertility cases are related to female endocrinological problems and costs around 1 billion Euro per year. Assisted Reproduction Techniques (ART) have made huge improvements on chances of infertile couples to have their own genetic babies. Since the birth of the first 'in vitro' child in 1978 (Kamel, 2013), ART have been widely used in fertility clinics. Assisted reproductive technology achieves pregnancy by using either in vitro fertilisation (IVF), intracytoplasmic sperm injection (ICSI), or other methods (Szmelskyj and Aquilina, 2014).

Systems biology is a complex approach to tackle an entire organism, instead of singling out fractions and trying to understand them. It came a long way from first mathematical model of cardiac cells leading to the implantable pacemaker, to the whole-cell mathematical model which is able to predict response to genetic mutations (Karr et al., 2012). It combines knowledge from biology to computer science, from engineering to physics aiming to predict how a system will react to a changing conditions and to develop solutions to the major healthcare problems of our time. One of which is to assist doctors in clinical practice by delivering clinical decision support systems.

The intention of this thesis is to apply systems biology to the problem of infertility. A solid amount of research has been done towards design a whole-body model such as Physiomodel (`www.physiomodel.org`) and Open Systems Pharmacology Suite (OSPS) (`https://github.com/open-systems-pharmacology`) that allows us to model also interaction with drugs, however, none of them closely deal with endocrinological problems thus, they do not fully covers the problem of infertility. A great deal of work was done specifically oriented on recreating the dynamics of reproductive hormones as model in (Röblitz et al., 2013) study. Such models have a high complexity and more than 100 parameters to be identified. Despite the ability to simulate concentration of hormones, the problem of identifing values for such a large amount of unknown parameters remains unresolved or highly complex.

Whereas models as (Röblitz et al., 2013) oriented on simulating the dynamics of multiple hormones such as Progesterone, Follicle-Stimulating Hormone, Luteinizing Hormone within normal cycle, this thesis oriented on establishing several models designed specifically for Estradiol concentration and follicle dynamics within stimulation treatment, which will be later on discussed in Section 2.

1.2 Motivation

Assisted Reproduction Techniques have increased chances for infertile couples, however, current success rates reach only 35% even in modern clinics. Therefore, the major goal of this thesis is to improve quality of fertility treatments, which consists of downregulation (or preparation) and stimulation phases. While downregulation aims at suppressing FSH, LH hormones, intention of stimulation phase is to obtain large number of mature follicles. Clinician follows up a patient response via measurements of hormone E2 and TV-US. We aim to develop models for both types of measurement.

The main inspiration for the first part of our research is (Bächler et al., 2014). The increase in the follicular surface area in species correlates linearly both with species mass and with the predicted increase in E2 concentration. This suggests that E2 grows linearly with the total surface of follicles. We aim to develop an E2 estimation model based on follicle sizes that, reducing E2 measurements during fertility treatments, will provide to physicians the same amount of information as they would have by measuring E2. As a matter of fact, E2 estimation from follicle measurements allows doctors to estimate E2 without waiting for the results of laboratory tests. Moreover, the ability to estimate E2 blood concentrations opens up an opportunity for healthcare at a distance, with the help of small devices which are available on the market (Sonaura (2016), Gerris and De Sutter (2010)). These devices allow patients to take TV-US by themselves at home and transmit results (via the Internet) to the doctor. Moreover, since less blood samples would be taken from a patient, it means less expenses, thus stimulation treatment would reduce it cost.

Our second aim is to develop a follicle model, which predicts future total and average surface of follicles within a TV-US, from a preceding TV-US measurement of follicle sizes and a drug dose during stimulation treatment. Stimulation treatment is a complicated process, where one of the crucial roles is played by the determination of the best day for ovulation induction. Clinician must decide this day based on a patient measurement and his/her professional experience. Moreover, high hormone doses administered to a patient could lead to dangerous treatment advers affects, such as Ovarian Hyper-Stimulation Syndrome. With the help of our follicle model a clinician could predict future dynamics of follicles, thus avoid a possibility of OHSS, as well as have more information to decide on ovulation induction day.

Our models, once reliable, could be integrated in Decision Support System that supports clinician during stimulation treatment.

1.3 Contribution

This thesis aims at reducing the number of measurements taken from a patient in order to increase patient comfort and reduce treatment costs. During a stimulation treatment mainly two types of measurements (see Figure 2.1) are performed by clinicians in order to monitor patient response to treatment, TV-US and E2 hormone concentrations. Both type of measurements are invasive and have to be taken every couple of days (depending on the protocol). In this thesis, we introduce 3 models for reducing both types of measurements taken from a patient. First, we present two families of models for E2 level blood estimation from the number and sizes of growing follicles during fertility treatments. In the first model, we exploit biological knowledge following a grey box approach. We assume that follicle contribution to E2 blood concentration depends linearly on the follicle granulosa layer surface. In

the second family of models, we split follicles into a certain number of classes, each of which contains follicles whose diameter is in a given range. Each follicle contributes to E2 blood concentration depending on the class it belongs to. This model is more general and is mainly a black box model, since it does not assume any specific relationship between follicle surface (or diameter) and E2 concentration.

Our piece-wise family of models has correlation coefficient higher than 0.7 for 14 patient groups out of 15. Moreover, 11 groups out of 15 have correlation coefficient more than 0.8. Both families of models reveal similar errors for an estimation error validated by bootstrap method. In most of patient groups 2-PL and 3-PL models do not outperform 1-PL significantly, nevertheless there are several exceptional groups.

Our third model is oriented to predict future surface of follicles, based on previous measurements and drug doses that were administered to a patient. This model is a piece-wise model predicting total and average surface of follicles and it does not assume any specific relationship between follicles. Our follicle model on average provides an error of 17.5% for a patient group. In 7 patient groups out of 9 the error is less than 20% (Table 4.1b). The error validated by bootstrap method on average gives error around 30%.

We observe that in both E2 and follicle models parameters are likely to be population as well as treatment dependent. Therefore, to take advantage of them, each clinic should fit parameter values by using data collected during treatments carried out in that clinic.

1.4 Summary

This thesis is organised as follows. We first give a brief description to a stimulation treatment (Figure 2.1), followed up by common techniques applied through the thesis in Section 2.1. The thesis contains two principal parts. Section 3 presents two type of models to estimate E2 concentration. One that assumes a linear dependency between E2 and granulosa layer in a follicle, Sect.3.2.2.1. Second that is mainly data driven, see Section 3.2.2.2. Section 4 describes a model for predicting the outcome of TV-US, with knowledge of previous measurement and the injected drug dose. Each part (Section 3, 4) contains experimental results. Finally, we discuss obtained results in Section 5 and close the topic in Conclusion Section 6.

Chapter 2

Context

This chapter provides a gentle introduction to the thesis topics and in particular to aspects of fertility (treatments). We first briefly describe the infertility treatment in Section 2.1 and then discuss common techniques which we use in our study (Section 2.2). Data used in our study is described in the following Section 2.3. Finally, we discuss follicle measurements errors, and thus intrinsic limitations to predictability of our models in Section 2.4.

2.1 Fertility Treatment

Fertility treatments consist of downregulation (or preparation) and stimulation phases. Downregulation aims at blocking the release of Follicle Stimulation Hormone (FSH) and Luteinizing Hormone (LH) responsible of proliferating activity of granulosa cells in follicles and, thus, stops the development of antral follicles. After patient responds to downregulation, clinicians start the stimulation phase, whose goal is to lead a large number of follicles to maturation and then collect oocytes for subsequent in vitro fertilisation (IVF) or intracytoplasmic sperm injection (ICSI).

Follicle maturation is achieved by administering high doses of follicle stimulating hormones (FSH or a mix of FSH and LH). In order to optimise drug administrations, the number and quality of collected oocytes, and to minimise risks of treatment adverse effects (most serious being Ovarian Hyper-Stimulation Syndrome (OHSS) (Mason et al., 1994)), patient treatment response is monitored during follicle development by a series of Trans-Vaginal Ultrasounds (TV-US), and measurements of Estradiol (E2) blood concentrations. Main clinical decisions during the stimulation phase, such as doses and timing of drug administrations, when to induce ovulation, and possibly to stop the treatment depend on the number and sizes of growing follicles and E2 blood concentrations.

The general structure of a stimulation phase is depicted in Figure 2.1. Some decisions are taken in advance (i.e., before starting the treatment) on the basis of external factors or measurements taken before the beginning of treatment. As shown in Figure 2.1, a treatment is a loop that ends when treatments goals are met, or when there is evidence that such goals cannot be attained any more. In this case, the treatment ends with a failure. During the treatment, some safety conditions, that guarantee patient health have to be always satisfied. If such conditions are violated, the treatment ends with a failure. However, if success goals are met with respect to safety conditions, treatment considered to be a success and it terminates the loop.

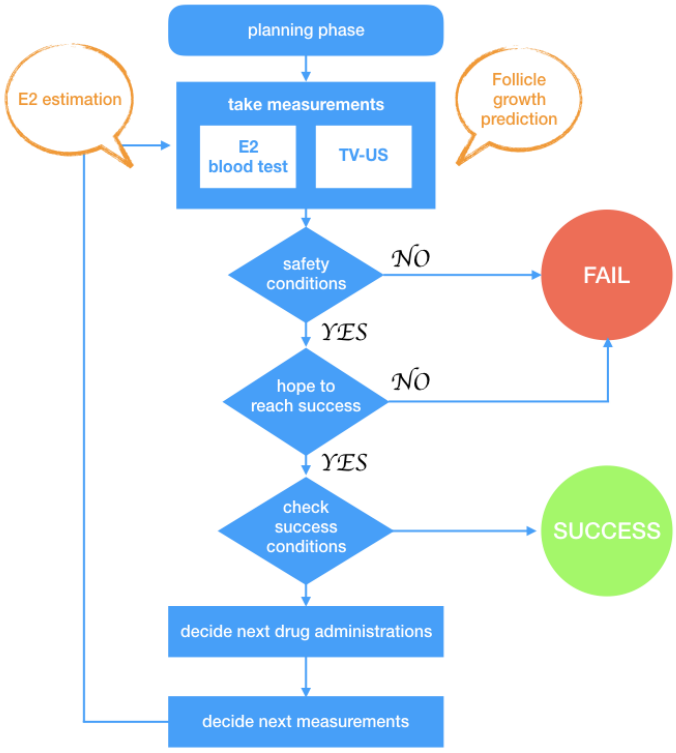

FIGURE 2.1: Stimulation phase general structure

During stimulation treatment a patient is at reask to obtain an Ovarian hyperstimulation syndrome (OHSS). To monitor this condition clinicians follow closely an E2 blood concentration. If E2 level of a patient reaches a certain threshold, clinician could adjust or terminate a treatment in order to avoid OHSS. The mitigation of risk to obtain OHSS is presented on Figure 2.1 by Safety Conditions.

Under stimulation treatment, many follicles grow and become mature oocytes, while the others undergo atresia. Stimulation treatment success (data collected inside the PAEON project at University Hospital Zurich) is defined by the presense of at least three mature follicles with a diameter ≥ 16 (mm).

For the convenience, follicles are usually classified accordingly to their diameter. We introduce a Follicle Profile (FP) definition with respect to the following diameter classes. Classes = <10, 10–11, 12–13, 14–15, 16–17, 18–19, $\geq$ 20>.

Definition 1. *The follicle profile of a medical case in a given day is a function defining the number of follicles within each diameter class for that medical case in that day.*

Definition 2. *We call a medical case one course of stimulation treatment for a patient.*

2.2 Techniques

This section summarizes a general schema for the model design and assessment. This schema applies both for E2 hormone concentration part (Section 3) and Ovarian Follicle dynamics part (Section 4). Both parts present mathematical models with unknown parameters to be identified. Thus, we first present grey box modelling approach, based on both biological knowledge and experimental data, in Section 2.2.1.1 and review follicle structure in Section 2.2.1.2, which is the foundation for both parts of this thesis. Second, we present an inclusion criteria for a patient in our

study. We then discuss parameter idendification problem (see Section 2.2.3), which we solve by building an optimization problem and solve it with AMPL tool. Next, we present an improvment to our models by introducing patient groups, based on external factors, such as health condition, injected stimulation drug, see Section 2.2.4. Finally, we discuss a method to validate our methodology by using leave-one-out bootstrap technique and obtaining bootstrap average error and standard deviation (see Section 2.2.5).

2.2.1 Modelling Approach and Assumptions

2.2.1.1 Grey Box Modelling

In our study, we use biological knowledge to design our parametric models and experimental data in order to identify parameters and to validate accuracy of model predictions. This corresponds to the so-called *grey box modelling* approach (Sohlberg, 1998). Assuming a linear dependency between follicle surface and E2 concentrations helps to keep the model simple (linear models are easy to define) and parameters, estimated on experimental data, can be easily interpreted by clinicians.

In contrast, white-box models use only one type of knowledge about a system – physical knowledge, while black-box models use only experimental data. Grey-box modelling has several advantages which are valuable for our mathematical model. Most importantly, we can use our prior biological knowledge about a system. Grey-box modelling mixes both kind of data – experimental data and biological. Given the fact that we indeed have both of this data type, it is reasonable to use grey-box modelling.

2.2.1.2 Follicle structure

Essentially, a follicle has an elliptical shape (Penzias et al., 1994) and a multilayered structure, as shown in Fig. 2.2. It consists of a fluid-filled antrum, and granulosa and theca layers. While SonoAVC software (Raine-Fenning et al., 2008) is an emerging approach to measure follicle volume (for example, to identify pathology or confirm normality of follicle), still, 2D manual measurement is the standard approach to measure follicles during stimulation treatments.

Retrospective data considered in our study only record one diameter per follicle. Therefore, in our study, we will assume follicles to have a spherical shape. The contribution of a follicle having diameter d to the overall E2 blood concentration is proportional to the surface of its granulosa layer (Bächler et al., 2014). In (Bächler et al., 2014), it is also shown that the thickness of the granulosa and theca layers can be considered constant across different patients and during follicle growth. Hence, the surface of the granulosa layer of a follicle having diameter d can be estimated as $S = \pi(d - 2t)^2$, where t is the thickness of theca layer, that we always set to $100\mu m$ (Bächler et al., 2014).

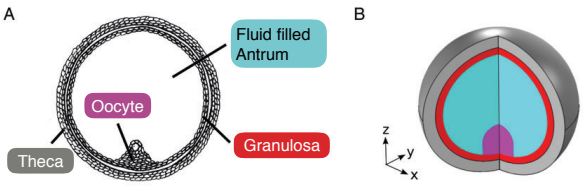

FIGURE 2.2: (Bächler et al., 2014) (**A**) A schematic 2D representation of an ovarian follicle. (**B**) A schematic view of the 3D computational domain for the follicle

2.2.2 Medical Case Log: Inclusion Criteria

The following criteria must be all satisfied by a medical case log in order for it to be included in our study:

1. log contains all information concerning external factors of a medical case, in order to be classified to a group
2. log refers to a medical case having exclusively one disease

A log is excluded if:

1. it refers to a medical case with the value of AFC equals to 1

2.2.3 Model Validation

Both parts of this thesis, E2 hormone concentration part (Section 3) and Ovarian Follicle dynamics part (Section 4), represent mathematical models with unknown parameters to be identified. Thus, it is our goal to identify them. We find values for such parameters by solving optimisation problems, in order to minimise the mismatch between model predictions and available measurements. In principle, all these parameters are medical case dependent and it would be desirable to identify such parameters for the medical case at hand. Unfortunately, finding individualised parameters would require several measurements in advance. By contrast, our aim here is to provide to physicians methods and tools to estimate E2 hormone concentration and Ovarian Follicle dynamics while reducing number measurements.

2.2.4 Improvement on Model Predictions: Patient Groups

As inter-patient parameters lead to unsatisfactory predictions, we introduce patient groups. Each group is identified by a set of *external factors*. In our research, we consider 4 external factors. The first is the Antral Follicle Count (AFC) that is how many antral follicles are present at the beginning of the cycle. AFC reflects woman fertility potential. The second indicates whether a woman is healthy or has some infertility causes. The third factor discriminates medical cases in which the stimulation is performed by administering FSH only and those in which a blend of FSH and LH is administered. The last factor corresponds to the clinician that performed TV-US. In our experimental results, this factor confirm that accuracy in taking measurements is crucial for data considered in parameter identification.

More precisely, we split medical cases according to the following criteria:

1. ***AFC, measured before stimulation treatment:*** we have considered 2 classes: Medium response, where $5 \leq$AFC<10 (MR) and Elevated, where $10 \leq$AFC<20 (ER). Although we have considered other level of response (Low Response ($2 \leq$AFC<5) and High (AFC≥ 20)), our retrospective data does not contain enough medical cases in these classes.

2. ***Health Condition:*** we considered 4 categories: healthy medical cases, (i.e. women without hormonal infertility causes), Endometriosis, Idiopathic (unknown reasons of infertility), and Other reasons.

3. ***Administered drug:*** we considered two categories of administered drugs - FSH only and drugs containing a combination of FSH, LH.

4. ***Measurement technique:*** in our data set, received from UZH, measurements were taken by three different people, thus, it introduces some variability to measurements and that is why person identification was chosen as one of the factors to split medical cases in groups.

For each response level r, each health condition h, each administered drug d, and each measurement technique t, we have a group $g = (r, h, d, t)$. By considering medical cases in a group g, we compute inter-patient group parameter values v_g^* finding, *for each group* g those values that minimise average errors between estimations and measurements.

2.2.5 Methodology Validation

Our final step is to validate our methodology, by using leave-one-out bootstrap technique and obtaining bootstrap average error and standard deviation (Hastie, Tibshirani, and Friedman, 2009). First, we give a notation to the bootstrap itself and afterwards show how it could be used.

General wish using bootrap is to get statistical accuracy of some quantity $S(Z)$, based on original dataset Z. Let's say $Z^{*1}, Z^{*2}, ..., Z^{*B}$ are randomly drawn datasets with replacement having same size as original. We do this B times, as it is shown in Figure 2.3.

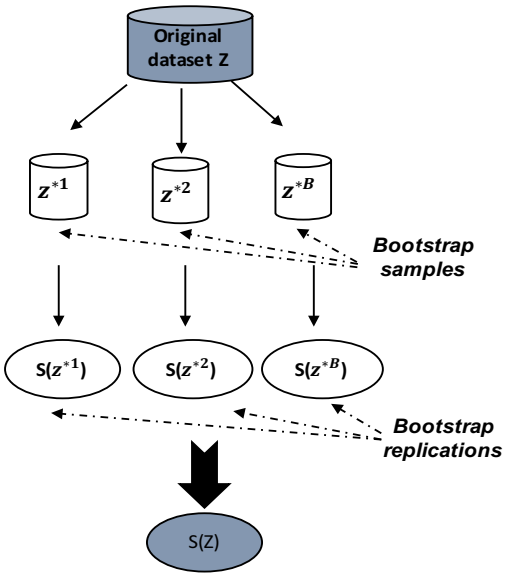

FIGURE 2.3: General view on bootstrap tool.

The quantity $S(Z)$ is computed from each bootstrap set and the values $S(Z^{*1})$, $S(Z^{*2})$, ..., $S(Z^{*B})$ are used to assess the statistical accuracy of $S(Z)$. Using simple bootsrap may lead to imprecise estimation, since training bootstrap samples and original testing set contain common samples. In order to overlap this drawback, for each observation we consider only bootstraps which does not contain observation

i (leave-one-out bootstrap). Thus, leave-one-out bootstrap estimate of prediction error is defined as in (3.4).

$$Err = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-E2_i}|}\sum_{b\in C^{-E2_i}} L(y_i, f^b(x_i)) \tag{2.1}$$

Where $f^b(x_i)$ is the predicted value at x_i, from the model fitted to the bth boot- strap dataset and C^{-i} is the set of indices of the bootstrap samples b that do not contain observation i. We either have to choose B large enough to ensure that all of the $|C^{-i}|$ are greater than zero, or we can just leave out the terms corresponding to $|C^{-i}|$ that are zero.

2.3 Dataset Statistics

This thesis was done in the framework of the PAEON (PAEON, 2016a) research project on eHealth and Virtual Physiological Human funded by EU FP7. Project documentation is available online (PAEON, 2016b).

In our study, we have considered one dataset containing data on stimulation treatments. Data is collected inside the PAEON project at University Hospital Zurich (UZH) and remains private due to the Project regulations. The dataset contans 624 patients, with overall number of medical cases 1037, see Table 2.1. As we stated in Section 2, one patient commonly goes under several rounds (medical case) of stimulation treatment. Figure 2.6 shows that 40% of patients have a course of stimulation treatment at least twice or more. Figure 2.7 shows a number of full observations (both follicle profile and E2 measurement were performed on the same day) that were performed on a patient, most common number of observations is 2. Our dataset also contains information about external factors, including diseases. More than 50% of patients have 1 disease (Figure 2.5). Time distance between medical cases associated to the same patient is shown on Figure 2.4.

Our study mainly cosists of two parts, E2 hormone concentration and ovarian follicle dynamics. In both parts (Section 3 and Section 4) we included patient groups having at least 10 medical cases.

The E2 hormone concentration model was included in the PAEON (PAEON, 2016a) research project.

TABLE 2.1: High-level statistics of the dataset.

Overall number of patients	624
Overall number of medical cases	1037
Overall number of observations	2019

2.4 Underestimation by TV-US

Usually in clinical practice follicles are measured manully using 2D image from ultrasound by taking either the mean of two largest diameters or by taking the largest. However, it is not always the case that follicle has a spherical shape, often follicles has an irregular shape. Thus, the use of a traditional 2D ultradound leads to high inaccuracy. Study in (Raine-Fenning et al., 2008) analysed 224 follicles within the

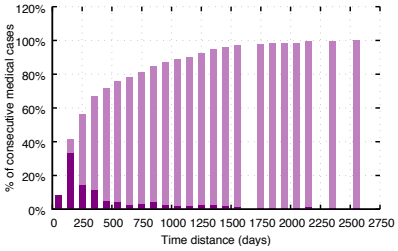

FIGURE 2.4: Time distance between consecutive medical cases associated to the same patient.

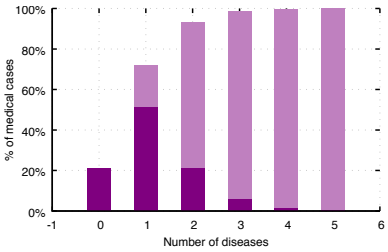

FIGURE 2.5: Number of diseases per medical case.

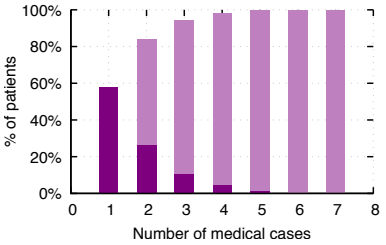

FIGURE 2.6: Number of medical cases per patient.

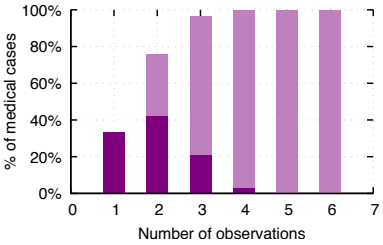

FIGURE 2.7: Number of observations per medical case.

volume range $0.4 - 16.2$ (cm^3). They compare four ways to obtain volume of follicles. First is an automatic way (using SonoAVC), second is to calculate volume using three diameters obtained by SonoAVC and the sphere formula. Third way is based on three diameters as well, however the numbers were obtained from 2D ultrasound. Last one is calculated using Virtual Organ Computer-aided AnaLysis (VOCAL). Third way had the highest error in case of a single diameter. To compare, they calculated mean and standard deviation of volume. While SonoAVC provided particularly close results (3.67 ± 2.51) to true follicle volume (3.70 ± 2.60), 2D with one diameter gained 4.40 ± 3.42. Even additional measurements of diameters (two/three diameters) contains high error.

Aditionally, study in (Rosendahl et al., 2010) compares ovarian volume estimated by the 2D TV-US with the ovarian volume measured after weigh of unilateral oophorectomy. This study included 66 women who had an ovary removed for cryopreservation of the ovarian cortex. It concluded that ovarian volume was severly underestimated by at least 27% from measurement obtained by 2D transvaginal ultrasound. This high error takes place due to the fact that 2D TV-US is based on mathematical model assuming that ovary has a shape of prolate ellipsode, yet in reality it is not always true.

This underestimation by TV-US is in agreement with our study. Our models tend to underestimate both hormone levels and follicle growth, as model parameters are fitted by using 2D measurement data. Precision of our model predictions are clearly intrinsically limited by precisions of measurements we use to fit model parameters.

2.5 Quantisation Task

Our models assume that the $\hat{E}2$ produced by each single follicle is proportional to the surface of its granulosa layer. However, quite often in standard clinical practice only information about a diameter class is available and not the exact diameter (in our data from obtained from University Hospital Zurich, only a 2 mm wide diameter class is known for each follicle). This makes the surface of the granulosa layer for each follicle an *uncertain* quantity.

We overcome this problem by observing that, when assigning each observed follicle to its diameter class, the human operator executed an instance of a *quantisation task*. As a consequence, relying on the typical assumptions made when dealing with quantisation tasks, our models assume that the actual measured diameter for follicle is drawn *uniformly at random* within its diameter class $[\hat{d} - \delta; \hat{d} + \delta]$ (with $\hat{d}$ being the mean diameter of such a class). Thus, the *expected* surface for the granulosa layer of follicle f, can be computed by instantiating to this case the standard formula for the expected value of a random quantity:

$$\tilde{S}_f^G = \int_{\hat{d}-\delta}^{\hat{d}+\delta} \pi(x - 2t)^2 \frac{1}{2\delta} dx \tag{2.2}$$

Before presenting our results on the accuracy of our E2 estimation model, in this section we show the range of values for error (see Section 2.2.5) of formula 3.4 that we can regard as *satisfactory*.

The surface of a granulosa layer is an *uncertain* quantity for two reasons:

1. Follicle diameters in our retrospective data have been obtained with a manual inspection of TV-US clinians, who derived some sort of mean diameter for each follicle. Such mean diameters are in no way average diameters in any geometrical sense. Also, they might be subject to errors, see Section 2.4.

2. For each follicle only information about its 2mm-wide diameter class is available, and not its measured mean diameter.

Hence, in order to correctly interpret the values of the bootstrap error of 3.4 when assessing performance of our estimator, it is crucial to understand what is the impact of such uncertainties on the final validation error value.

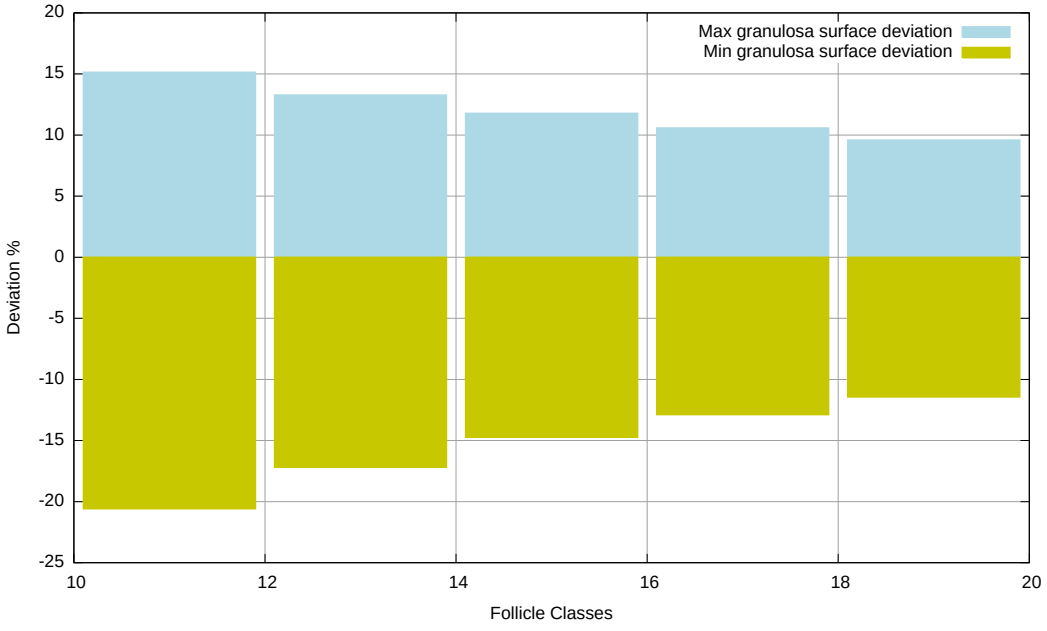

FIGURE 2.8: Bounds of V G ratios for follicles of each diameter class.

Figure 2.8 shows the extreme variations of the granulosa surface of a follicle of each diameter class with respect to the expected value as defined in 2.2. In particular, for each follicle diameter class, the figure shows the ratios $\frac{S^G_{max}}{\bar{S}^G}$, $\frac{S^G_{min}}{\bar{S}^G}$ between the maximum (minimum) value of the surface of the granulosa layer for a follicle of that diameter class and the expected granulosa layer surface as defined in 2.2.

Given that E2 estimation model is linear to the overall follicular granulosa surface, the bounds above also define a range of uncertainty in the estimated E2 value which cannot be neglected.

As a result, when we assess performance of our estimation model on our data, we will consider bootstrap values within the bounds shown in 2.8 as *satisfactory*, because such error values could be *fully justified* in terms of the intrinsic uncertainties in the input (about exact mean follicle diameters).

Chapter 3

E2 Hormone Concentration

This chapter is aimed to present our E2 estimation model. We first discuss our motivation and related work for the study in Section 3.1. Then we present two family of models for E2 estimation: *piece-wise* models in Section 3.2.2.1 (that assume a linear dependency between E2 and surface of granulosa layer) and *step-wise* models in Section 3.2.2.2 that are mainly data driven. We then compare them in Section 3.2.2.3. Finally, we present our experimental results in Section 3.3.

3.1 Introduction

Despite a still open debate (Kwan et al., 2014; Orvieto et al., 2008), E2 blood level is an important factor considered in clinical decisions during fertility treatments (Malhotra, 2015; Mittal et al., 2014; Var et al., 2011). Even though some authors (Vandekerckhove et al., 2014; D'Angelo et al., 2004) claim that ultrasounds are enough and E2 measurements are needed only if OHSS risks are high due to other factors, the majority of authors as well as clinicians (Kwan et al., 2014; Malhotra, 2015) strongly support monitoring *both* follicle growth and E2 levels as a good clinical practice.

3.1.1 Motivation

It is well known that, during follicle maturation, E2 is synthesised mainly by granulosa cells surrounding oocytes in ovarian follicles (Mason et al., 1994). Our main goal here is to design a *quantitative* model that faithfully estimates E2 from the number of growing follicles and their sizes.

The main inspiration for our research is (Bächler et al., 2014), where it is shown that while the size of mature oocytes is similar across different mammalian species, the size of ovarian follicles differs greatly. The increase in the follicular surface area in larger species correlates linearly both with species mass and with the predicted increase in E2 concentration. This suggests that E2 grows linearly with the total surface of follicles.

With respect to (Bächler et al., 2014), we aim at developing an E2 estimation model based on follicle sizes that, reducing E2 measurements during fertility treatments, will provide to physicians the same amount of information as they would measure E2. The benefits of such a model would be remarkable in clinical practice in terms of patient comfort, treatment costs, and logistic.

As a matter of fact, E2 estimation from follicle measurements allows doctors to estimate E2 without waiting for the results of laboratory tests. Moreover, the ability to estimate E2 blood concentrations opens up an opportunity for healthcare at a distance, with the help of small devices which are available on the market (Sonaura (2016), Gerris and De Sutter (2010)). These devices allow patients to take TV-US by themselves at home and transmit results (via the Internet) to the doctor.

3.1.2 Contribution

In this work, we introduce and evaluate two families of models for E2 level blood estimation from the number and sizes of growing follicles during fertility treatments.

First, we consider *piece-wise affine* models (*n-Piece-wise Legs* (*n*-PL)). In this family of models, we assume that follicle contribution to E2 blood concentration depends *linearly* on the follicle granulosa layer surface. Moreover, we let the slope change during follicle maturation, according to the observation that E2 secretion rate changes at different follicle maturation stages. These models depend on the number n of times we allow the slope changes (legs), the n breakpoints $\gamma_1, \ldots, \gamma_n$ in which slope changes, and on the $n+1$ slopes $\beta_0, \beta_1, \ldots, \beta_n$.

Second, we consider *step-wise* models. In this family of models, we split follicles into k classes, each of which contains follicles whose diameter is in a given range $[a_k, b_k]$. Each follicle contributes to E2 blood concentration depending on the class it belongs to. These models depend on the parameter k (number of classes) and on k parameters $\lambda_1, \ldots, \lambda_k$, where λ_i models the contribution of a follicle in class i to E2 blood concentration. These models generalise the one considered in (Franco et al., 1993), where only 3 classes of follicles were considered.

All these models are *parametric*: we fit parameter values by solving optimisation problems, finding those values minimising the error of E2 estimations with respect to real E2 measurements in retrospective data. Unfortunately parameters turn out to be highly medical case dependent: inter-patient parameters lead to unsatisfactory model predictions. To obtain more reliable results, we split medical cases into *groups* and fit parameter values for each group separately. Groups are defined by medical case external factors (health condition and number of antral follicles at the beginning of the stimulation phase etc.) and by treatment properties (administered drugs during stimulation, follicle measurement techniques etc.).

We observe that model parameters are likely to be *population* as well as *medical cases* dependent. Therefore, to take advantage of them, each clinic should fit parameter values by using data collected during treatments carried out in that clinic.

In comparison to (Franco et al., 1993) model, our N-PL family of models is more precise in E2 estimation. While (Franco et al., 1993) model has correlation coefficient 0.7, all 15 patient groups have correlation higher (or equal) than 0.7. Moreover, 11 groups out of 15 have correlation coefficient more than 0.8. All correlation coefficients were calculated on leading N-PL model (see Table 3.2) in each group. As for the estimation error validated by bootstrap method, both families of models reveal similar errors, see Figures 3.3 and 3.4. In most of patient groups 2-PL and 3-PL models do not outperform 1-PL significantly.

3.1.3 Related Work

Work presented by (Bächler et al., 2014) shows first of all correlation between surface area in a follicle, both with species mass and E2 concentration. It is worth pointing out that (Bächler et al., 2014) research studies natural cycles, while our research focuses on stimulated cycles during fertility treatment. The main difference between natural cycles and stimulated cycles is that, under stimulation, many follicles grow and become mature oocytes, whereas in a natural cycle usually only one (occasionally two) reaches maturity while the others undergo atresia. Accordingly, during stimulation, E2 levels are much higher than in a natural cycle. With respect to this work, we plan to investigate connection between E2 level and follicles under high doses of hormones.

To the best of our knowledge, the study closest to our investigation is (Franco et al., 1993). They divide follicles into just 3 groups with respect to their diameter: <14mm, in between 15 and 17mm, and >17mm on ovulation induction day and devise a relationship between E2 and follicle number and sizes. With respect to (Franco et al., 1993), we were more precise on follicle measurements and monitored the relationship between E2 and follicles number and sizes during the entire stimulation treatment. In comparison with (Franco et al., 1993) study gaining correlation coefficient of 0.7 between measurements of E2 levels and E2 estimations, our model is more precise and has correlation coefficient always higher (or equal in only one group) then 0.7. Among them we gain more then 0.80 correlation coefficient between measurements of E2 levels and E2 estimations in more then half of patient groups. This point is essential for our main practical objective, i.e. to reduce E2 measurements during stimulation, while keeping the same information for the doctor.

Quite large amount of research has been done in terms of prediction success rate of becoming pregnant with help of In Vitro Fertilization procedures. Main idea of these studies is to predict whether patient will or will not get pregnant based on oocyte quality, embryo quality, level of E2 and many other factors. These studies are mostly oriented to establish relationship between pregnancy rate and factors on oocyte retrieval day, while our study suggests a model to improve treatment quality during entire stimulation phase of fertiliy treatment, by estimating E2 consentration while oocytes are not yet developed. The general outcome of studies using statistical analysis (Orvieto et al., 2007), (Var et al., 2011), (Mittal et al., 2014), is that the ratio E2/number of oocytes is a good marker, proving the relevance of E2 levels in fertility treatments. Besides them, a number of studies used machine learning techniques to predict the same success rate, like (Kim and Jung, 2003) by carrying out Bayesian network-based analysis detects that age and stimulants like hCG, FSH, LH, Clomiphene, Parlodel and GnRH play the key role in pregnancy of an infertility patient. Also (Passmore et al., 2003) build decision trees with accuracy of 67.4 %, in order to predict patient success of becoming pregnant using IVF procedures.

TABLE 3.1: Related work summery

Study	Study investigates how E2 grows	Study suggests to monitor E2	Study suggests to monitor only in cases with OHSS threat	Study suggests to monitor TV-US	Study investigates if IVF success rate is influenced by E2 concentration	Study investigates factors influencing IVF success rate	Study include monitoring OHSS	Stimulated cycles(•)/Natural cycles(∗)
(Kwan et al., 2014)			•	•			•	•/
(Aboulghar, 2003)			•				•	•/
(Al-Hussaini, 2012)			•	•			•	•/
(Malhotra, 2015)		•		•			•	•/
(Gerris and De Sutter, 2010)			•	•				•/
(Bächler et al., 2014)	•	•		•				/∗
(Franco et al., 1993)		•		•				•/
(Orvieto et al., 2008)				•			•	•/
(Vandekerckhove et al., 2014)			•	•				•/
(D'Angelo et al., 2004)			•				•	•/
(Papanikolaou et al., 2006)			•	•			•	•/
(Orvieto et al., 2007)		•		•	•	•		•/
(Var et al., 2011)		•		•	•	•	•	•/
(Mittal et al., 2014)		•		•	•	•		•/
(Kim and Jung, 2003)						•		•/
(Passmore et al., 2003)					•	•		•/

3.2 Methods

In this section, we outline our mathematical models for estimating Estradiol (E2) blood concentration from follicle measurements during the stimulation phase of a fertility treatment together with our approach to model assessment.

We present in Sect. 3.2.2 the two (families of) models that we consider in our work. Our parameter identification technique is presented in Sect. 3.2.3. Finally, in Sect. 3.2.4, we present our approach to assess model prediction accuracy.

3.2.1 Modelling Approach

In our study, we use biological knowledge to design our parametric models and experimental data in order to identify parameters, see Section 2.2.1.1. Assuming a linear dependency between follicle surface and E2 concentrations helps to keep the model simple (linear models are easy to define) and parameters, estimated on experimental data, can be easily interpreted by clinicians.

3.2.2 Parametric Models

In this section, we present our 2 families of models: *piece-wise* models in Sect. 3.2.2.1 (that assume a linear dependency) and *stepw-ise* models in Sect. 3.2.2.2 that are mainly data driven. Finally, we compare them in Sect. 3.2.2.3.

3.2.2.1 E2 Piece-wise Linear Models

It is well known that E2 is synthesised mainly by ovarian granulosa cells (Mason et al., 1994). The main assumption in the design of piece-wise models is that E2 concentration grows linearly with respect to the surface of granulosa layer as established by (Bächler et al., 2014). This study, however, considered natural cycles (across different mammalian species), whereas our research goal is to estimate E2 from the surface of granulosa layer during the stimulation phase of a fertility treatment. Indeed, under ovarian stimulation, many follicles grow and become mature oocytes, whereas in a natural human cycle usually only one (occasionally two) reaches maturity, while the others undergo atresia. Therefore, E2 growth in natural and stimulated cycles can greatly differ.

We generalise this idea by considering models in which E2 depends not just linearly from granulosa layer in a follicle, but it also depends on the stage of follicle maturation. This leads us to consider parametric *n-leg piece-wise linear* (n-PL) models, in which E2 concentration piece-wise linear function with $n-1$ break points where slope changes, as shown in Fig. 3.1.

Our parametric piece-wise model depends on parameters α, β, γ where $\alpha \in \mathbb{R}_{\geq 0}$, $\beta = \langle \beta_1 \dots \beta_n \rangle \in \mathbb{R}^n$ and $\gamma = \langle \gamma_1 \dots \gamma_{n-1} \rangle \in \mathbb{R}^{n-1}$ with $0 \leq \gamma_1 \leq \gamma_2 \leq \dots \leq \gamma_{n-1}$. To simplify formulas, we add to the vector γ values $\gamma_0 = 0$ and $\gamma_n = \infty$. These parameters represent an offset α, E2 growth rates β_i (with respect to follicle surface), and break points γ_i where the ratio between E2 and follicle surface changes (see Figure 3.1, where $\hat{E2}$ is the estimated level of E2).

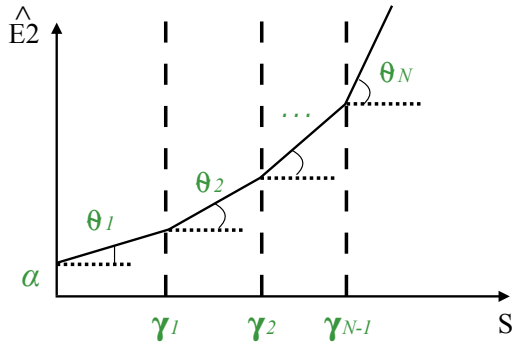

FIGURE 3.1: E2 mathematical model as an n-leg Piece-wise Linear Function connecting surface of granulosa layer (S) in a follicle and E2.

Given a follicle of surface S, such that $S \in [\gamma_k, \gamma_{k+1}]$, we model its contribution to the total E2 concentration as:

$$E2(S) = \beta_{k+1}(S - \gamma_k) + \sum_{j=1}^{k} \beta_j(\gamma_j - \gamma_{j-1}) \tag{3.1}$$

Total E2 is then simply the sum of the contribution of all follicles. If, in a given measurement m, there are f follicles having surfaces $S_1, \ldots, S_f$, we put $E2(m) = \alpha + \sum_{i=1}^{f} E2(S_i)$.

Experimentally, in Sect. 3.3, we will show that in most patient groups 2-PL and 3-PL models do not outperform 1-PL model significantly. However, in some cases 2-PL model gives the lowest error in contrast to 1-PL, which confirms our hypothesis that follicle influence on E2 concentration depends on its stage of maturation. In this case, we call first stage *primary* (associated to the parameter $\beta_1 = \beta_p$), corresponding to early stage of follicle development (follicle size less than γ) having low impact on E2 concentration. We call *maturation* (associated to the parameter $\beta_2 = \beta_m$) corresponding to a mature follicle stage (size greater than γ), when the follicle has a much stronger influence on E2 concentration.

As a result, we obtain family of piece-wise linear models, depending on $2n$ unknown parameters.

3.2.2.2 E2 Step-wise Models

Also in this family of models, we assume that E2 concentration is proportional to the sum of E2 secreted by each follicle. Moreover, here we assume follicles uniformly distributed, and that the E2 secreted by each follicle depends on an unknown smooth function f of its diameter.

As a consequence, having f follicles of diameters $d_1 \leq \cdots \leq d_f$ revealed in a measurement m, we have $\hat{E2}(m) = \lambda_0 + \sum_{i=1}^{f} f(d_i)$, where λ_0 is E2 not secreted by follicles.

If we split the range of possible values of d_i into k intervals $[a_j, b_j]$ ($j \in \{1, \ldots, k\}$), and we let $\mu_j = (a_j + b_j)/2$, we have, by Taylor's Theorem, that $f(d_i) = f(\mu_j) + f'(\xi_j)(d_i - \mu_j)$, where $d_i \in [a_j, b_j]$. If the width $b_j - a_j$ of the interval $[a_j, b_j]$ tends to 0, and $d_i \in [a_j, b_j]$, then $f(d_i)$ tends to $f(\mu_j)$. Thus, assuming f exists, then all follicles within same diameter range $[a_j, b_j]$ contribute to E2 concentration approximatively for the same quantity $f(\mu_j) = \lambda_j$.

Therefore, we have that $\hat{E2}(m) = \lambda_0 + \sum_{i=1}^{f} f(d_i) \approx \lambda_0 + \sum_{j=1}^{k} n_j f(\mu_j)$, where n_j is the number of follicles having diameter in $[a_j, b_j]$. When the width of intervals tends to 0, then the error tends to zero. Accordingly, if we have arbitrary precise measurements, we can have arbitrary precise models for E2 estimation.

As a result, we obtain a family of models, depending on $(k+1)$ unknown parameters, $\lambda_0, \lambda_1, \ldots, \lambda_k$ to be identified, in which:

$$\hat{E2}(m_i) = \lambda_0 + \sum_{j=1}^{k} n_j \lambda_j \tag{3.2}$$

3.2.2.3 Comparing Piece-wise Linear to Step-wise Models

On the one hand, step-wise models are more general and are mainly *black box* models, since they do not assume any specific relationship between follicle surface (or diameter) and E2 concentration. On the other hand, step-wise models prediction ability is limited by the measurement precision in the parameter identification phase. As a matter of fact, once the granularity of the model has been chosen, the prediction of the model is constrained by that choice.

For example, if follicles are measured with a precision of $2mm$, at it is the case in our retrospective data, and parameter $\lambda_1, \ldots, \lambda_k$ are identified by using such data, the model will treat as equals all follicles that belong to a class of dimension of $2mm$, regardless of more precise measurements.

By contrast, exploiting biological knowledge, (in our case, linear dependency between follicle surface and its contribution to E2 total concentration), following the *grey box* modelling approach of piece-wise linear models, one can benefit from more precise measurements, by interpolating contribution of each follicle. For example, even if model parameters are identified by using a resolution of $2mm$, they can take advantage from accurate measurements (for example with a precision of 1 or $0.1mm$).

In both families of models, measurement accuracy during model parameter identification is crucial. As it will be shown in Sect. 3.3, in our experimental study, prediction ability of our models is strongly influenced by the measurement technique (that in our retrospective data is related to the person (nurse, operator) that perform (take) measurements).

3.2.3 Parameter Identification

In Sect. 3.2.2, we have described two families of models: piece-wise linear (Sect. 3.2.2.1) and step-wise (Sect. 3.2.2.2) models. Both of these models depend on parameters that have to be identified. Piece-wise linear models depend on $2n$ parameters, where n is the number of legs, that is the number of points in which the linear dependency between E2 and follicle size can change slope. Step-wise models depend on $k+1$ parameters, where k is the number of classes in which we classify follicles.

We find values for such parameters by solving optimisation problems, in order to minimise the mismatch between model predictions and available measurements (Sect. 3.2.3.1).

3.2.3.1 Parameters Identification

Let T be a set of t medical cases. For each medical case i, we have a set $M = \{M_{i,1}, \ldots, M_{i,m_i}\}$ of m_i measurements taken at different days during the stimulation phase of the fertility treatment. Each measurement $M_{i,j}$ consists of a value $E2_{i,j}$

(E2 blood concentration) and a set $D_{i,j} = \{d_{1,i,j}, \ldots, d_{f_{i,j},i,j}\}$ of $f_{i,j}$ follicle diameters revealed by the doctor performing TV-US.

Let p stand both for the tuple α, β, γ (parameters of piece-wise linear models) and the tuple λ (parameters of step-wise models). We denote with $\hat{E2}^v(D_{i,j})$ E2 estimations given by the model in which parameters p have been instantiated with the tuple of values v. Our aim is to find a tuple v^* of values that minimises the average relative error between model predictions $\hat{E2}^v(D_{i,j})$ with respect to the measured value $E2_{i,j}$. Formally, we find v^* as:

$$v^* = argmin \sqrt{\frac{1}{M} \sum_{i=1}^{t} \sum_{j=1}^{m_i} \left(\frac{\hat{E2}^v(D_{i,j}) - E2_{i,j}}{E2_{i,j}} \right)^2} \tag{3.3}$$

where $M = \sum_{i=1}^{t} m_i$

Thus, we are facing a problem of quadratic optimisation. The objective function described in Equation (3.3) is the average error between model estimations $\hat{E2}$ and the $E2$ concentrations recorded in measurements taken in t medical cases.

3.2.3.2 Relative and Absolute Errors

In our study we choose to optimize relative error for both parameter identification (Section 3.2.3.1) as well as for evaluation (Section 3.2.4) of our model.

We find parameter values optimizing both absolute (absolute optimizer) and relative errors (relative optimizer), as we discussed in Section 3.2.3.1. This leads to diverse results. Then we compare them by calculating relative percentage error of measurements. In Figure 3.2 we show the distinction for one patient group. As we can see, optimizer build for an absolute error gives very high error on relatively small measurements of E2 (this area marked with red on the figure). Yet on a relatively high values of E2 it provides similar error to the error obtained by relative optimizer (this area as well marked with red on the figure).

Figure 3.2 compares errors only for one patient group. Appendix C contains same graphs for the rest of patient groups and Table 3.4 contains absolute error values for all patient groups, considered in our study. We also present distribution graphs for those experiments in Appendix A

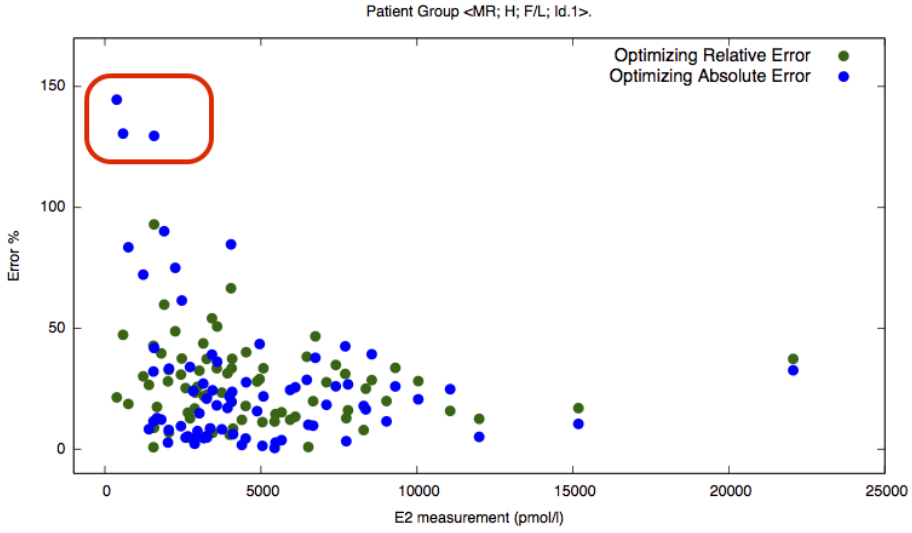

FIGURE 3.2: Shows relative percentage error for each measurement of E2 in the <MR, H, F/L, Id.1> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the optimal model in NPL family, see Table 3.2, where optimal model in each family of models is coloured with purple.

3.2.4 Model Evaluation Approach

We have discussed the leave-one-out bootstrap technique in Section 2.2.5, where estimate of prediction error is defined as in 2.1. In this Section we present an estimate of prediction error for the E2 model as in 3.4.

$$Err = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-E2_i}|}\sum_{b\in C^{-E2_i}}\left|\frac{\hat{E2}_i(b) - E2_i}{E2_i}\right| \tag{3.4}$$

Where C^{-E2_i} is the set of indices of the bootstrap samples b that do not contain observation $E2_i$. We either have to choose B large enough to ensure that all of the $|C^{-i}|$ are greater than zero, or we can just leave out the terms corresponding to $|C^{-i}|$ that are zero. We chose B to be 100 in our experiments.

3.3 Results

In this section we are facing three matters of concern. Our main objective addresses the issue of models validation (A) and comparison of E2 model estimation performances among them. In order to validate models, we chose to calculate the error minimizing the difference between E2 measured and estimated, taking into account all measurements from a patient group. We called this error *deterministic error*. As expected, experiments showed that the higher number of legs is involved in piece-wise model, the lower is deterministic error. The same principle stands for the step-wise models, the more narrow are follicle classes, the lower is deterministic error. Our second issue is to compare performances of our models to (Franco et al., 1993) study (B). We evaluated correlation coefficient between E2 measurements and E2 model estimations, and in all of the patient groups, we obtain coefficients higher then 0.7, as it was gained by (Franco et al., 1993). Finally, we validate our methodology (C), by applying leave-one-out bootstrap technique and obtaining *bootstrap average error*.

3.3.1 Experimental Setting

We have considered several settings for our families of models. Both families, step-wise as well as piece-wise, have multiple settings, which led to corresponding number of experiments. To be more precise, we have run them with the following settings - step-wise model with $K = 3$ (as in (Franco et al., 1993)), $K = 6$, $K = 12$ and piece-wise with 1-leg, 2-legs, 3-legs.

First, we have run experiments, for the same step-wise model as (Franco et al., 1993) with $K = 3$ is to compare performance of our models to it. Second we run the step-wise model with $K = 6$ (6-SW model), implying that $K + 1$ parameters have to be identified, see Section 3.2.2.2. Assuming uniform intervals and considering 24 mm as the maximum diameter of a follicle, this restriction means, that follicles within intervals of 4 mm will be treated as equal and will have only one parameter. Second, model with $K = 12$ (12-SW model) is the finer model that it makes sense to

consider, since our retrospective data used 2 mm precision while performing TV-US. Experiments have shown that 6-SW model may be used as an E2 estimation model and 12-SW model do not significantly outperform it.

Since step-wise models confirm linear connection between E2 concentration and follicle diameter, our next step is to run experiments for the N-PL models. Primarily, in our interest is to check whether 1-PL, or simply linear model, is the estimation model to proceed with, meaning 2-PL and 3-PL do not significantly outperform it. As experiments have shown, in most of the patient groups this turns out to be true, however two exceptions took place. Both of them are patient groups having medium response level, as one of external factors. The fact that 2-PL model outperforms 1-PL in a few groups, suggests that E2 concentration depends on the stage of follicle maturation.

As we discussed in Section 3.2.3.2, optimizer build for an absolute error gives very high error on relatively small measurements of E2. Besides, under such levels there is no evidence that the contribution of a follicle to E2 production is dominant. Thus, we chose to consider E2 measurements higher than 1000 (pmol/l) in our experiments.

To summarize, we consider 6 different models for each patient group (see Table 3.2). As a consequence, we compute for each patient group inter-patient error (see Sections 2.2.4 and 3.2.3.1), as well as correlation coefficient.

3.3.2 Model Validation (A), (B)

Both families of models described in Sections 3.2.2.1 and 3.2.2.2 are parameter dependent. Parameter values are identified by minimizing the error between E2 measurement and E2 estimation, using AMPL tool to solve the optimization problem in Equation (3.3). Before validating the error with the help of leave-one-out bootstrap (Hastie, Tibshirani, and Friedman, 2009) method, we first find the parameter values calculated on all measurements in a current patient group (A). We also calculate correlation coefficients in all groups, using deterministic parameter values, which allow us to compare our results to (Franco et al., 1993) study (B).

On average among patient groups deterministic error using 1-PL model is around 26.6%. Two groups suffer from high error values. Group <MR; Healthy FSH/LH; Id.2> with 35.9% for the deterministic error, <ER; Healthy; FSH/LH; Id.3> with 39.1% and group <MR; Other; FSH/LH; Id.3> with 33.8%. Two out of three groups have medium level of response, additionally, measurements for <ER; Healthy; FSH/LH; Id.3> and <MR; Other; FSH/LH; Id.3> were performed by same person, Id.3. As experiments showed in group <MR; Other; FSH/LH; Id.3> 2-PL model outperform 1-PL, while in two other groups it does not sustain. Two groups <MR; Healthy FSH/LH; Id.2> and <ER; Healthy; FSH/LH; Id.3> have an unpredictable behaviour and bootstrap error showed to be more then 34%. This supports importance of accurate measuremenrs and unpredictability of several groups. On another side, there are several groups where 2-PL model outperform 1-PL. For example, in patient groups <MR; Other; FSH/LH; Id.2> and <MR; Idiopathic; FSH/LH; Id.1> 2-PL outperform by around 5%.

As well as piece-wise family of models, step-wise family also provides low deterministic error values. Step-wise model with $K = 3$, on average through patient groups, has value around 26.3%. The same three patient groups <MR; Healthy FSH/LH; Id.2>, <MR; Other; FSH/LH; Id.3>, <ER; Healthy; FSH/LH; Id.3> have an unpredicted behaviour with deterministic error values 39.7%, 32.7%, 41.0% respectively. Also, step-wise models have patient groups with significantly lower

optimization error with increased number of steps.In patient group <MR; Other; FSH/LH; Id.2» 6-SW outperform almost by 5%. On top of this, several more groups has 6-SW model outperform 3-SW by around 6% for example group <MR; Other; FSH/LH; Id.3>.

In comparison to (Franco et al., 1993) where autors have found a correlation coefficient of 0.7, both of our families of models provide higher (or equal) correlation values between E2 estimations and E2 measurements in all groups. We obtain average correlation coefficient of 0.86 using 1-PL model and it does not significantly change in 2-PL and 3-PL. Only two patient groups have value in between 0.7 and 0.8. The rest of fourteen groups have correlation coefficient higher then 0.8. At maximum, correlation coeffient is 0.95 in patient groups <MR; Idiopathic; FSH/LH; Id.1> and <MR; Healthy; FSH/LH; Id.1>. At minimum 0.7 in patient group <ER; Healthy; FSH/LH; Id.3> with unpredictable behaviour. Correlation values computed with step-wise model estimation are close to those obtained with piece-wise model estimation. Using $3-SW$ model same patient group <ER; Healthy; FSH/LH; Id.3> with unpredictable behaviour has coefficient lower then 0.7, only two group within 0.7 and 0.8, and rest groups have coefficient higher then 0.8. At maximum, correlation coeffient is 0.95 in <MR; Idiopathic; FSH/LH; Id.1> and <ER; Healthy; FSH/LH; Id.1>. At minimum 0.67 in <ER; Healthy; FSH/LH; Id.3>. Model $6-SW$ provides quite similar correlation coefficients, with the exception of four groups, where it outperforms $3-SW$ model. Using 12-SW model we obtained correlation coefficient higher then 0.9 in 7 patient groups and value in between 0.7 and 0.8 in only one group, all the rest patient groups have values higher then 0.8.

On average, we gain correlation coefficient of 0.86 using 1-PL and 0.87 using 2-PL model, which are significantly greater than 0.7 gained by (Franco et al., 1993) study. However, some groups have an unpredictable behaviour, most seemingly because of the lack of precise measurements, since measurements for two groups with unpredictable bevaviour were performed by the same person. This fact indicate the need for a prospective study, with careful measurements to fix models parameters and considering usage of three-dimensional ultrasound imaging (Raine-Fenning et al., 2008).

See Patient Group description in Table 3.5.

3.3.3 Methodology Validation (C)

Each model setting resulted in *boostrap average error*, (see Sections 2.2.5, 3.2.4) which is obtained by applying leave-one-out bootstrap technique, with number of random samples equals to 100, see (Hastie, Tibshirani, and Friedman, 2009). Figures 3.3, 3.4 show all models and average errors corresponding to them, both for medium and elevated response groups. Despite the fact that we could estimate level of E2 concentration, it is worth pointing out that there are two groups with boostrap error higher then 35%. On another hand, there are three patient groups with boostrap average error less than 20%. All the rest groups have error in between 20% and 30%.

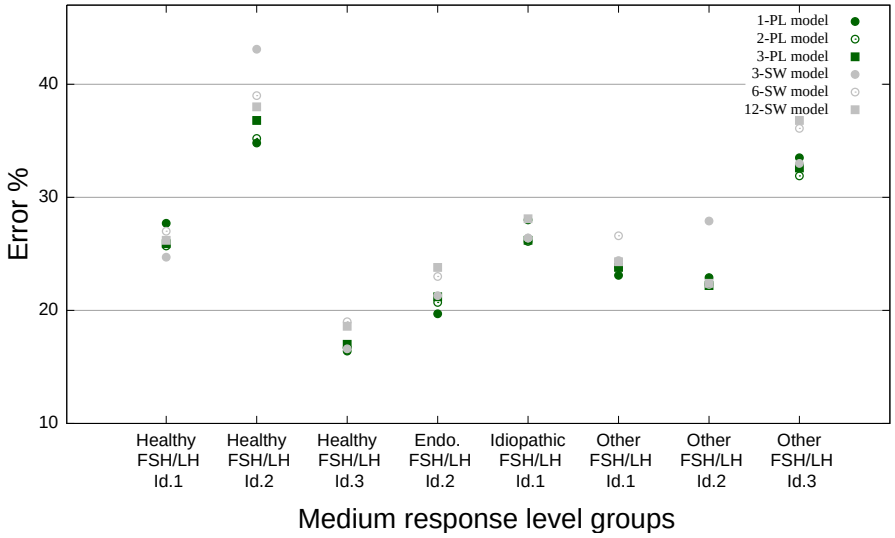

FIGURE 3.3: Estimated bootstrap error for medium level of response groups. For piece-wise linear models: 1-PL, 2-PL, 3-PL, as well as for step-wise models: splitting wad done by 4 mm and by 2 mm.

On average boostrap error for medium response groups is around 25% both for 1-PL model and 2-PL, see Figure 3.3. Two-legs model does not outperform 1-PL significantly, with at maximum boostrap error for medium response groups 34.8% and minimum 19.7%. Same goes for the 3-PL model, which has 25.0% on average for boostrap error and minimal at 17.0% and maximum with 36.8%. Experiments showed that medium reponse section has two groups where 2-PL model outperform 1-PL. Measurements for both of the <Idiopathic; FSH/LH; Id.1>, <Other; FSH/LH; Id.3>. As for step-wise family of models, bootstrap error is quite close to piece-wise models.

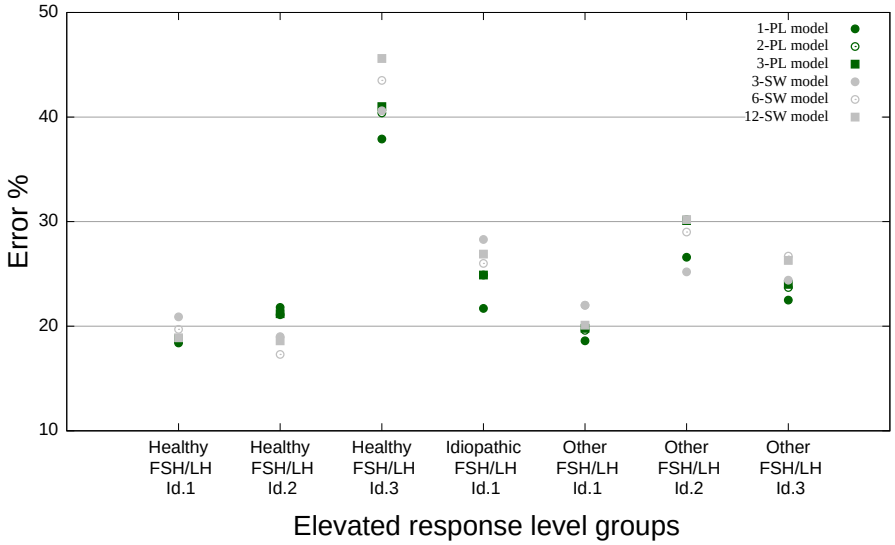

FIGURE 3.4: Estimated bootstrap error for elevated level of response groups. For piece-wise linear models: 1-PL, 2-PL, 3-PL, as well as for step-wise models: splitting wad done by 4 mm and by 2 mm.

Within elevated response section bootstrap error on average is 23% with minimum error 18.4% in group <Healthy; FSH/LH; Id.1> and maximum 37.9% in <Healthy; FSH/LH; Id.3> for 1-PL. Models with 2 and 3 legs have similar values to 1-PL. Another interesting fact is that in some groups, for example <ER; Healthy; FSH/LH;

Id.1> on Figure 3.4, piece-wise models give lower bootstrap error in comparison to step-wise models. While, original error calculated on all measurements, before using bootstrap method, for step-wise models is lower then in piece-wise models - as expected to be.

A clear tendency of overfitting in $12 - SW$ model is seen through the experiments, as expected to be. Another tendency is that some patient groups has a lower error in $6 - SW$ model in comparison to $3 - SW$ models.

3.3.4 Comparing E2 Estimations

As we pointed out in Section 3.3.3, there are several groups where 2-PL model outperform 1-PL, thus, this supports our hypothesis stating that in some patient groups, E2 depends not just linearly from granulosa layer in a follicle, but it also depends on the stage of follicle maturation. Figures 3.5 and 3.6 show E2 estimation points for two patient groups <MR, Other, FSH/LH, Id.2>, <ER, Healthy, FSH/LH, Id.2>, using different piece-wise models. Estimation points, calculated using 2-PL model, turned out to be more close to measurements than estimation points, calculated using 1-PL. It is supported by both Figures 3.5 and 3.6, where green points are measurements, blue are estimation points from 1-PL and purple estimation points from 2-PL.

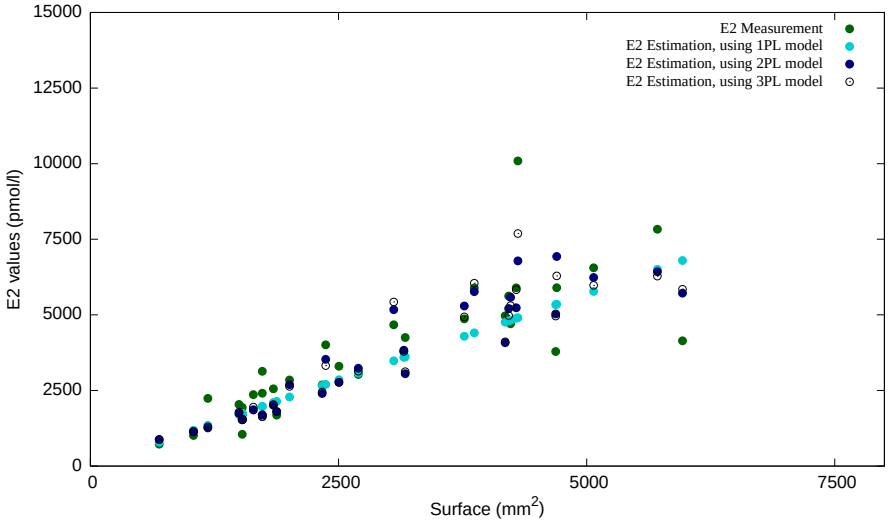

FIGURE 3.5: Estimation of E2, for <MR, Other, FSH/LH, Id.2> group, using 1-PL, 2-PL, and 3-PL models. Group <MR, Other, FSH/LH, Id.2> is on the several groups, where 2-PL model outperform 1-PL. Measurements taken from this group are coloured with green.

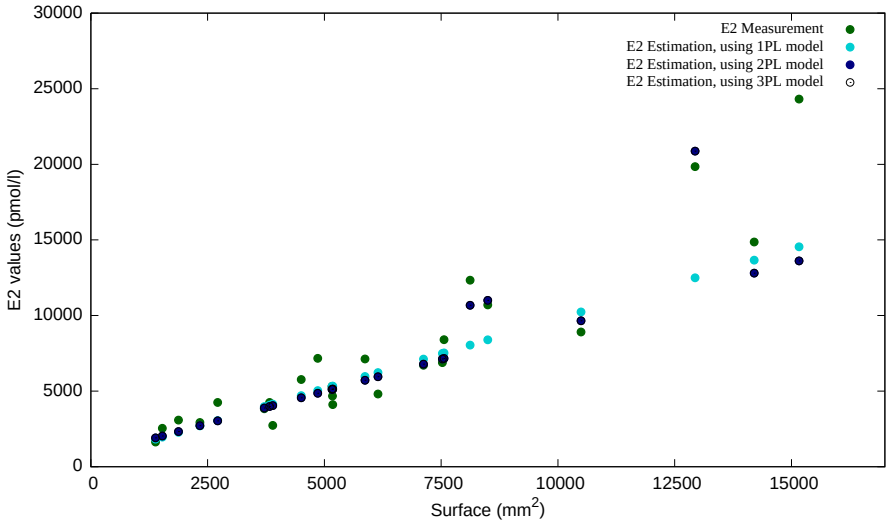

FIGURE 3.6: Estimation of E2, for <ER, Healthy, FSH/LH, Id.2> group, using 1-PL, 2-PL, and 3-PL models. Group <ER, Healthy, FSH/LH, Id.2> is on the several groups, where 2-PL model outperform 1-PL. Measurements taken from this group are coloured with green.

If we compare Figures 3.5 or 3.6 to 3.7, one could clearly identify the difference between them. Figure 3.7 shows measurement points and estimation points, as well as Figures 3.5 and 3.6, but there is no essential variation between estimation points obtained from different piece-wise models, i.e. estimation points are quite close to each other. Meaning that in group <MR, Healthy, FSH/LH, Id.3> stage of follicle maturation plays no role and E2 depends simply linearly from granulosa layer in a follicle.

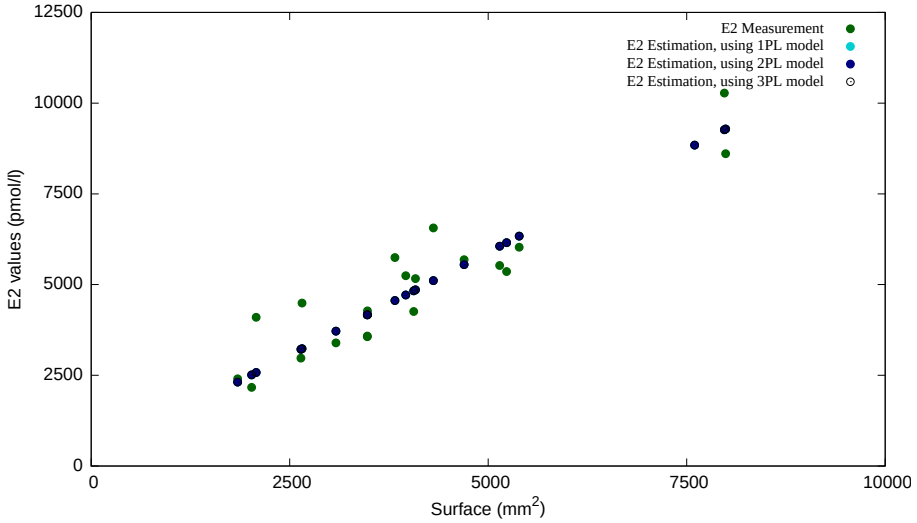

FIGURE 3.7: Estimation of E2, for <MR, Healthy, FSH/LH, Id.3> group, using 1-PL, 2-PL, and 3-PL models. In this group 2-PL model does not significantly outperform 1-PL.

3.3.5 E2 Estimation Service

Within the framework of the PAEON project, we have developed an Estradiol Estimation (E2E) software service. This service provides an E2 estimation based on

external factors, selected by the user, see Figure 3.9. If such combination of external factors is not valid, estimator works if user provides at least 2 full measurement (E2 and FP), see Figure 3.8. Based on 2 input measurements, service optimizes the patient-specific parameters and provides a user with an estimation of E2. In case of valid external factors user only needs to provide a FP and obtain an E2 estimation from the service.

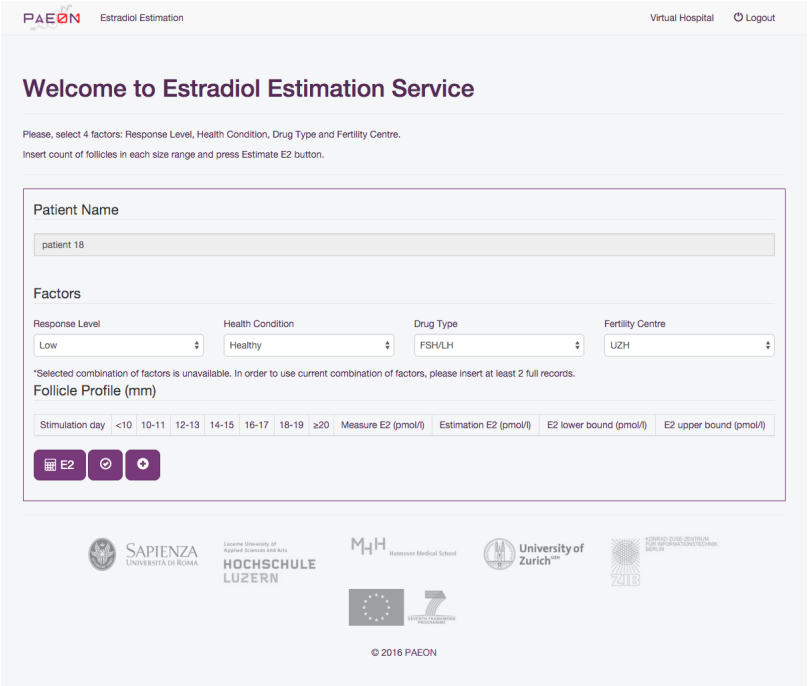

FIGURE 3.8: Estradiol Estimation software service. First estimation is possible if current combination of external factors is available or if user provides 2 full measurements (E2 and FP).

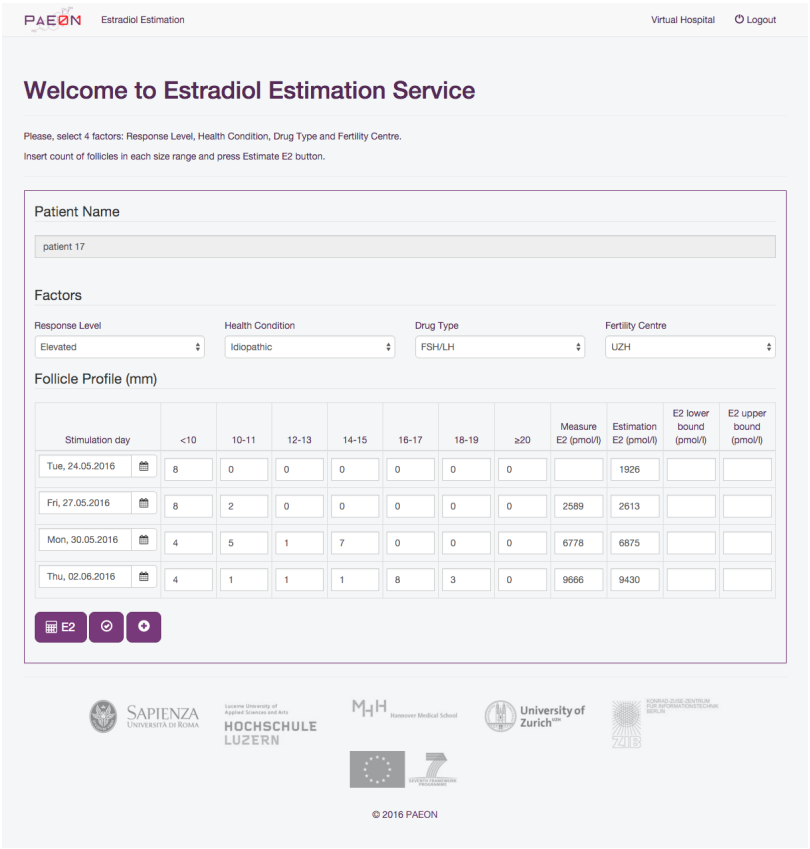

FIGURE 3.9: Estradiol Estimation software service.

Gr. Id	1-PL model (*)	2-PL model	3-PL model	3-SW model (Franco et al., 1993)	6-SW model	12-SW model
	c.c. / d.e.	c.c. / d.e.	c.c. / d.e.	c.c. / d.e.	c.c. / d.e.	c.c. / d.e.
1	0.90 / 30.8	0.90 / 27.8	0.90 / 27.8	0.91 / 27.7	0.90 / 29.8	0.91 / 26.9
2	0.80 / 35.9	0.80 / 33.3	0.80 / 33.1	0.72 / 39.7	0.79 / 34.6	0.85 / 31.1
3	0.87 / 17.5	0.87 / 17.5	0.87 / 17.5	0.89 / 17.5	0.88 / 16.5	0.89 / 15.8
4	0.78 / 22.9	0.79 / 22.8	0.79 / 22.8	0.79 / 22.3	0.79 / 21.8	0.80 / 21.6
5	0.95 / 26.2	0.95 / 23.0	0.96 / 22.6	0.95 / 22.4	0.95 / 21.8	0.96 / 21.4
6	0.84 / 27.2	0.84 / 26.5	0.84 / 26.5	0.83 / 27.9	0.82 / 28.1	0.84 / 25.6
7	0.92 / 26.8	0.96 / 22.0	0.97 / 21.5	0.94 / 27.0	0.97 / 22.5	0.97 / 20.9
8	0.87 / 33.8	0.89 / 29.6	0.89 / 29.6	0.88 / 32.7	0.87 / 29.4	0.88 / 29.2
9	0.95 / 20.5	0.95 / 19.7	0.95 / 19.7	0.95 / 22.1	0.95 / 20.2	0.96 / 18.6
10	0.92 / 23.7	0.91 / 20.7	0.91 / 20.7	0.94 / 18.7	0.93 / 17.0	0.93 / 17.0
11	0.70 / 39.1	0.70 / 39.1	0.70 / 39.1	0.67 / 41.0	0.70 / 38.8	0.7 / 38.8
12	0.86 / 23.7	0.88 / 22.2	0.88 / 22.1	0.84 / 24.8	0.89 / 22.2	0.89 / 21.8
13	0.91 / 22.3	0.91 / 20.8	0.92 / 20.7	0.90 / 24.9	0.94 / 23.1	0.92 / 20.3
14	0.80 / 25.9	0.82 / 24.1	0.82 / 24.1	0.83 / 22.5	0.82 / 22.7	0.84 / 22.3
15	0.94 / 23.9	0.95 / 23.6	0.95 / 23.6	0.94 / 24.2	0.94 / 24.0	0.95 / 22.0

TABLE 3.2: Relative Error Values (%) on the Training sets. Models coloured with green, give the lowest error for a current patient group. Patient groups coloured with yellow, have lower or equal error value for 1-PL model in comparison to 3-SW model (Franco et al., 1993). See Patient Group description in Table 3.5.

Gr. Id	1-PL model	2-PL model	3-PL model	3-SW model (Franco et al., 1993)	6-SW model	12-SW model
1	27.7	25.7	25.9	24.7	27.0	26.2
2	34.8	35.2	36.8	43.1	39.0	38.0
3	16.4	16.7	17.0	16.6	19.0	18.6
4	19.7	20.7	21.2	21.3	23.0	23.8
5	28.0	26.1	26.2	26.4	26.4	28.1
6	23.1	23.9	23.8	24.4	26.6	24.3
7	22.9	22.3	22.2	27.9	22.3	22.4
8	33.5	31.9	32.6	33.0	36.1	36.8
9	18.4	18.6	18.8	20.9	19.7	18.9
10	21.8	21.1	21.2	19.0	17.3	18.6
11	37.9	40.4	41.0	40.6	43.5	45.6
12	21.7	24.9	24.9	28.3	26.0	26.9
13	18.6	19.6	19.8	22.0	22.0	20.1
14	26.6	30.2	30.1	25.2	29.0	30.2
15	22.5	23.7	24.0	24.4	26.7	26.3

TABLE 3.3: Validation Error Values. Patient groups coloured with yellow, have lower or equal error value for 1-PL model in comparison to 3-SW model (Franco et al., 1993). Models coloured with green, give the lowest error for a current patient group. See Patient Group description in Table 3.5.

Gr. Id	1-PL model	2-PL model	3-PL model	3-SW model (Franco et al., 1993)	6-SW model	12-SW model
1	1521.2	1443.2	1442.4	1360.1	1444.4	1357.4
2	1682.6	1666.2	1666.1	1837.7	1682.9	1480.5
3	1369.2	1356.4	1356.4	1303.7	1316.0	1256.9
4	1872.2	1803.4	1803.4	1768.4	1800.4	1752.0
5	1359.2	1240.4	1230.2	1233.4	1188.7	1156.4
6	2346.5	2344.7	2344.6	2391.6	2412.6	2266.6
7	1437.0	802.9	766.2	1024.1	833.0	724.6
8	1253.7	1102.1	1102.1	1140.0	1218.9	1067.8
9	1205.6	1152.0	1152.0	1154.5	1219.7	1103.5
10	2155.2	1877.6	1871.4	1716.6	1715.2	1715.2
11	2723.3	2679.8	2679.8	2751.1	2583.1	2579.1
12	2008.1	1777.7	1718.5	1920.2	1612.0	1612.0
13	1591.1	1472.1	1472.0	1617.3	1540.1	1457.4
14	1591.0	1481.4	1481.0	1472.0	1399.6	1321.9
15	1859.4	1777.4	1777.4	1934.5	1753.9	1601.6

TABLE 3.4: Absolute Error Values (pmol/l) on the Training sets. See Patient Group description in Table 3.5.

Patient Group ID	Patient Group
1	<Medium Response; Healthy; FSH/LH; Id.1>
2	<Medium Response; Healthy; FSH/LH; Id.2>
3	<Medium Response; Healthy; FSH/LH; Id.3>
4	<Medium Response; Endometriosis; FSH/LH; Id.1>
5	<Medium Response; Idiopathic; FSH/LH; Id.1>
6	<Medium Response; Other; FSH/LH; Id.1>
7	<Medium Response; Other; FSH/LH; Id.2>
8	<Medium Response; Other; FSH/LH; Id.3>
9	<Elevated Response; Healthy; FSH/LH; Id.1>
10	<Elevated Response; Healthy; FSH/LH; Id.2>
11	<Elevated Response; Healthy; FSH/LH; Id.3>
12	<Elevated Response; Idiopathic; FSH/LH; Id.1>
13	<Elevated Response; Other; FSH/LH; Id.1>
14	<Elevated Response; Other; FSH/LH; Id.2>
15	<Elevated Response; Other; FSH/LH; Id.3>

TABLE 3.5: Contains Patient Group description and a corresponding to it Patient Group ID.

Chapter 4

Ovarian Follicle Dynamics

This chapter presents an ovarian follicle dynamics under influence of stimulation treatment. It starts with a brief introduction, by discussing motivation and contribution of our work (Section 4.1). We then talk more deeply about our model and modelling approach in Section 4.2, 4.2.3. Evaluation of our model is presented in Section 4.2.5. We close this section by presenting an experimental results obtained from our model (Section 4.3).

4.1 Introduction

Key part in stimulation treatment is played by iterative measurements of TV-US (Kwan et al., 2014; Malhotra, 2015). It assists a clinician on a number matters of concern. Primary, measurements of TV-US help a clinician to decide next appointment for a patient. Additionaly, a clinician could follow up a patient responde to stimulation drugs through them. Based on follicle dynamics he/she could adjust drug dose, in order to regulate response to a treatment.

4.1.1 Motivation

During stimulation treatment clinician schedules a number of appointments for a patient. The goal of this is to optimally choose a day for ovulation induction, as well as to monitor the risks of treatment adverse affects such as OHSS - Ovarian hyperstimulation syndrome. The timing of future appointment clinician decides for each patient individually, based on past measurements and his/her professional experience. Our follicle dynamics model could support clinician on this decision. A clinician could use our model to predict the FP dynamics and depending on it to take a decision on how soon a patient should pay him/her a visit.

Within the framework of the PAEON (PAEON, 2016a) project, one of the tools developed was a TDSS (Treatment Decision Support System). The TDSS was oriented to first of all support clinician decisions as well as to help medical students (during their residency) to learn abount stimulation treatments. In order to assist on questions regarding infertility treatments, support system should be able to capture a follicle dynamics under influence of stimulation treatment. Our follicle model could be integrated in TDSS and improve it.

While E2 estimation model opens up an opportunity for healthcare at a distance, integration of it with follicle dynamics model opens up even a brighter prospective. Clinician who has at hand only a FP measurement, could predict a future dynamics of it using our follicle model and after obtain an E2 estimation. Suchwise, clinicial will have a prediction of a full measurement. The benefits of such combination would be remarkable in clinical practice in terms of patient comfort, treatment costs, and logistic.

4.1.2 Contribution

In this work, we introduce and evaluate a model for predicting total and average surface of Follicle Profile (FP), from a preceding measurement of FP and a drug doses during stimulation treatment.

Since during stimulation treatment high doses of FSH or FSH/LH are administered to a patient on a daily basis, it is reasonable to assume that depending on a treatment day growth speed of follicles is different. We consider a follicle dynamics model as a piece-wise model in which not only follicle growth depends on a linear combination of total and average surface, but it also depends on the phase of a stimulation treatment. Thus, we split stimulation treatment into 7 phases in between day 0 till day 20, allowing us to evaluate total and average surface of FP with respect to phase. Commonly, available data for a patient group has no more than 2 phases.

Our follicle model is parametric: we fit parameter values by solving optimisation problems, finding those values minimising the error of total and average surface predictions with respect to real FP measurements in retrospective data. Unfortunately parameters turn out to be highly patient-dependent as we saw for the E2 model parameters, see Section 3.1.2.

We observe that model parameters are likely to be population as well as treatment dependent. Therefore, to take advantage of them, each clinic should fit parameter values by using data collected during treatments carried out in that clinic.

4.1.3 Related Work

A great deal of work was done specifically oriented on recreating the dynamics of the human menstrual cycle (Röblitz et al., 2013), (Egli, Leeners, and Kruger, 2010). Unfortunately those models have no flexibility, have high complexity and frequently designed for some specific aims. As (Röblitz et al., 2013) for simulating the downregulation part of a fertility treatment (see Section 2.1) and the model in (Egli, Leeners, and Kruger, 2010) for analysing prolactin patterns.

Some follicle models are desined with a sole purpose. Studies in (Baerwald, Adams, and Pierson, 2003), (Panza, Wright, and Selgrade, 2016) specifically designed to analyze if folliculogenesis (maturation of the ovarian follicle) occurs in a wave-like fashion. Others, due to the fact that multiple follicle are under development at each ovarian cycle, studies in (Clément and Monniaux, 2013) and (Conover et al., 2001) were designed to determine the mechanisms underlying follicle selection.

Plenty of research is oriented towards developing a follicle dynamics model, there are even models where you could trace growth of each follicle individually (Clément et al., 2013), (Echenim et al., 2005). Additionally, studies (Echenim et al., 2005), (Soboleva et al., 2000) allow an exogenous dose of FSH, allowing to trace influence of it.

Despite an amount of research that has been done towards follicle model development, it is worth pointing out that most of this work is oriented towards natural cycles, whereas our research is focused solely on stimulated cycles.

Furthermore (Panza, Wright, and Selgrade, 2016), (Echenim et al., 2005) models have a high complexity and contain at minimum 25 (up to 120) parameters to be identified. Some studies take into account measurements obtained from patients (Panza, Wright, and Selgrade, 2016), while others validate models solely to the purpose of recreation time evolutions for the model species that are compatible with the law of biology.

In spite of the ability to simulate dynamics of an individual follicle, the problem of identifing values for such a large amount of unknown parameters remains unresolved or highly complex. We took a less ambitious approach by building a discrete-time model and gain parameters values based on measurements obtained from patients.

Additionally, we do not concern ourselves with obtaining dynamics of an individual follicle. As we discussed in Section 3.2.2.2 we assume that E2 concentration is proportional to the sum of E2 secreted by each follicle in the step-wise family of models, thus we aim to obtain total and average surface of follicles measured during a TV-US.

4.2 Methods

In this section, we present a mathematical model for predicting total and average surface of FP, based on a previous measurement of FP at time t and an injected stimulation drugs (FSH/LH or FSH) during the stimulation phase of a fertility treatment.

We start by describing our assumptions in Section 4.2.1. We then present in Section 4.2.3 the follicle growth model that we consider in our work. Next we present parameter identification technique in Section 4.2.4. Followed up by the idea of splitting a stimulation treatment into phases with respect to a stimulation day. Finally, in Section 4.2.5, we present our approach to assess model prediction accuracy.

4.2.1 Modelling Approach and Assumptions

In contrast to follicle models (Clément et al., 2013), (Echenim et al., 2005), (Panza, Wright, and Selgrade, 2016), we do not aim at a model defined by differential equations that describe integrations between all biologocal components involved in follicle growth. Instead we offer a follicle dynamics model as a piece-wise model in which not only follicle growth depends on a linear combination of total and average surface, but it also depends on the phase of a stimulation treatment.

This leads us to consider parametric k piece-wise linear (k-PL) model, with $(n+1)\cdot n\cdot k$ unknown parameters to be identified, where n is the number of species that we aim to predict. We consider n as the model size and k number of treatment phases. A model setting offered by us (see Section Results 4.3) has $n = 2$ (total and average surface). If number of treatment phases is equal to 2 (as it often is), then we have 18 parameters to be identified. Additionally, we consider model with $n = 3$, however as it will be shown in Results section, our follicle model could predict dynamics of follicles using size $n = 2$.

Knowing that FP surface grows with time and depends on FSH, we assume no more factors have an impact on a follicle growth. Thus, our approach could predict future value of total and average surface of FP, by identifing parameters for the model.

4.2.2 Treatment phase

During stimulation treatment medical case i obtains drugs on a daily basis. The effect of injections accumulates during those days and influences follicle growth differently in dependence of a treatment day. Thus, parameter values could divers depending on a day. Thereby, we split treatment duration into phases.

Let us define a treatment *phase*. Phase ρ is defined by an *initial* treatment day Ts, an *end* treatment day Tf and a set of medical cases T^{ρ}. Both on the initial and on the

end treatment days there must be at least one medical case with measurements of both E2 and TV-US (for model taking into account E2 measurements) and there must be at least one medical case with only measurement of TV-US on the initial and on the end treatment days for the model disregards E2. Phase ρ exists only if the initial and the end treatment days are defined.

Let us define a set of medical cases T^{ρ}. This is a set of medical cases whose measurements of E2 and TV-US happend in between the *initial* treatment day Ts and the *end* treatment day Tf, $Ts < Tf$.

$$T^{\rho} = \left\{ i \in T \;\middle|\; \begin{array}{l} \exists D_{i,j}.T_{i,j}^{(u)} = Ts, j \in [1, n_i]; \\ \exists D_{i,j}.T_{i,j}^{(u)} = Tf, j \in [1, n_i]; \\ \exists E_{i,j}.T_{i,j}^{(E)} = Ts, j \in [1, z_i]; \\ \exists E_{i,j}.T_{i,j}^{(E)} = Tf, j \in [1, z_i]; \end{array} \right\}$$

For each medical case i in a T^{ρ}, there is a sequence of TV-US measurements $D_i = [D_{i,1}, \ldots, D_{i,n_i}]$, where $D_{i,j}$ is basically a sequence of $f_{i,j}$ follicle diameters, $D_{i,j} = [d_{i,j,1}, \ldots, d_{i,j,f_{i,j}}]$, taken at days $T_i^{(u)} = [T_{i,1}^{(u)}, \ldots, T_{i,n_i}^{(u)}]$. Also, for each medical case i, there is a sequence of E2 measurements $E_i = [\, E_{i,1}, \ldots, E_{i,z_i} \,]$ taken at days $T^{(E)} = [T_{i,1}^{(E)}, \ldots, T_{i,z_i}^{(E)}]$. We should point that measurements of E2 and TV-US do not necessary occur on the same treatment day, thus z_i is not necessary equal to n_i.

- for each sequence $D_{i,j}$ we define a corresponding sequence of follicle surfaces $S_{i,j} = [S_{i,j,1}, \ldots, S_{i,j,f_{i,j}}]$, where $S_{i,j,k} = 4\pi(d_{i,j,k} \,/\, 2)^2$, $k \in [1, f_{i,j}]$.
- for each sequence $D_{i,j}$ we define a value of total surface, which is a sum of surfaces belonging to $S_{i,j}$. $Surface_{i,j} = \sum_{k=1}^{f_{i,j}} S_{i,j,k}$
- for each sequence $D_{i,j}$ we define an average surface as $Avg_{i,j} = \frac{1}{f_{i,j}} Surface_{i,j}$.

We may regard sequences as set when convenient.

4.2.3 Parametric Model

In our study we offer a model predicting growth of ovarian follicles under influence of stimulation treatment, see equation 4.1. More precisely our model predict growth of total surface of a follicle profile and average surface.

$$\dot{x}(t) = f(x(t), u(t)) = Ax(t) + Bu(t) \tag{4.1}$$

where, $A = (a_{ij}) \in \mathbb{R}^{n,n}$ is an unknown parameter matrix, $B = (b_1(t), b_2(t), \ldots, b_n(t))^T$ is a parameter vector. Using Euler method we can approximate the solution of the ODE (4.1), as follows.

$$x(t+1) = x(t) + Tf(x(t), u(t)) \tag{4.2}$$

where $x(t) = (Surface(t), Avg(t))^T$ and $u(t)$ is an input drug given at time t, and n is the model size.

Since, each treatment phase ρ depends on unknown parameters, we are facing an optimization problem with $(n+1) \cdot n \cdot k$ parameter values to be identified. We choose to minimise a relative error that optimizes the distance between measurements of TV-US.

4.2.4 Parameter Identification

In Sect. 3.2.2, we have described a model for predicting FP at next clinical appointment. Our model depends on parameters that have to be identified. We find values for such parameters by solving optimisation problems, in order to minimise the mismatch between model predictions and available measurements (Section 2.2.3).

4.2.4.1 Optimizing Average Error

Let γ stand both for the A, B parameters, see (4.1). We denote $y^{(k)}(t,\gamma)$ as a prediction (based on parameters γ) for a variable $x^{(k)}(t)$ and $\hat{x}^{(k)}(t)$ is a measurement of variable $x^{(k)}(t)$, $k \in [1,n]$.

Parameters γ have been instantiated with the tuple of values v. Our aim is to find a tuple v^* of values that minimises the average relative error between model predictions $y^{(k)}(t,\gamma)$ with respect to the measured value $x^{(k)}(t)$.

$$E_{RMS}(i,k) = \sqrt{\frac{1}{\left|T_i^{(k)}\right|}\sum_{t=1}^{\left|T_i^{(k)}\right|}\left(\frac{y_i^{(k)}(t,\gamma)-\hat{x}_i^{(k)}(t)}{\hat{x}_i^{(k)}(t)}\right)^2} \tag{4.3}$$

$$E_{AVG} = \frac{1}{n}\sum_{k=1}^{n} E_{RMS}(k) \tag{4.4}$$

4.2.4.2 Optimizing Error by Element

$$E_{RMS}(k) = \frac{1}{P}\sum_{i=1}^{P} E_{RMS}(i,k), k \in [1,n]; \tag{4.5}$$

4.2.5 Model Evaluation Approach

Our final step is to validate our methodology, by using leave-one-out bootstrap technique and obtaining bootstrap average error, see Section 2.2.5 and Equation 2.1. In this Section we present an estimate of prediction error for the follicle dynamics model as in 4.7 and in 4.9.

$$E_{RMS}(i,k,b) = \sqrt{\frac{1}{\left|T_i^{(k)}\right|}\sum_{t=1}^{\left|T_i^{(k)}\right|}\left(\frac{y_{i,b}^{(k)}(t,\gamma_b)-\hat{x}_i^{(k)}(t)}{\hat{x}_i^{(k)}(t)}\right)^2} \tag{4.6}$$

We denote $y_{i,b}^{(k)}(t,\gamma_b)$ as a prediction for a variable $x_i^{(k)}(t)$, based on parameters γ_b obtained from bootstrap b at time t, for a medical case i, $k \in [1,n]$. Root Mean Square error for a medical case i on element k, based on bootstrap b, is denoted by $E_{RMS}(i,k,b)$.

$$Err(k) = \frac{1}{|P|}\sum_{i=1}^{P}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} E_{RMS}(i,k,b) \tag{4.7}$$

$$St.dev(k) = \sqrt{\frac{1}{|P|}\sum_{i=1}^{P}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}}\left(E_{RMS}(i,k,b)-Err(k)\right)^2} \tag{4.8}$$

We optimize two types of errors - average and by element, see Sections 4.2.4.1, 4.2.4.2. Thus, bootstrap error for an element k is defined as in (4.7) and standard deviation in (4.8). Where P is a set of medical cases, C^{-i} is the set of indices of the bootstrap samples b that do not contain observations for medical case i.

Average bootstrap error on all n elements is an average error defined for each element k seperatly, see 4.7. It is defined in (4.9) and standard deviation in (4.10).

$$\begin{aligned} Err_{AVG} = \frac{1}{|P|}\sum_{i=1}^{P}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}}\frac{1}{n}E_{RMS}(i,k,b) = \\ \frac{1}{n}\sum_{k=1}^{n}\frac{1}{|P|}\sum_{i=1}^{P}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}}E_{RMS}(i,k,b) = \frac{1}{n}\sum_{k=1}^{n}Err(k) \end{aligned} \tag{4.9}$$

$$St.dev_{AVG} = \sqrt{\frac{1}{|P|}\sum_{i=1}^{P}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}}\left(\frac{1}{n}\sum_{k=1}^{n}E_{RMS}(i,k,b) - Err_{AVG}\right)^2} \tag{4.10}$$

As a result, we validate our model with respect to average error and by each element k seperatly. Thus, to each patient group used in our study will corresponde four types of values. First validating the model by minimizing average error, see Equations (4.7, 4.8) and by element, see Equations (4.9, 4.10).

4.3 Results

In this section, we are facing four matters of concern. First, our goal is to validate our model with respect to a patient (A), including all patient medical cases. In other words, one patient commonly goes several times under stimulation treatment. We call *medical case* a course of stimulation treatment for a patient. In order to do so, we choose to calculate the error minimizing the difference between measurements of species and estimations, taking into account all medical cases from a patient. We call this type of error - *patient-specific*. As expected, experiments showed that patient-specific error has low values, which testifies the correctness of our model.

Our second major goal is to validate our model with respect to patient groups (B.1). We point out that patient-specific specific predictions (A) have been performed on all available 147 patients who have more than 1 medical case, regardless if a patient belongs to a patient group (in agreement to her external factors) or not. While validatation of the model with respect to a patient group (B.1) (let us call them *inter-patient group parameters*) is performed for each group seperatly, solely on patients with external factors same as a patient group. We later validate our methodology with respect to a patient group (B.2), by applying leave-one-out bootstrap technique and obtaining bootstrap average error and standard deviation.

The number of available medical cases for a given patient is $n > 1$. As a last step, we investigate whether using $(n-1)$ previous meadical cases of a patient, we can predict patient behaviour on the n^{th} last available medical case (C). In order to do so, we minimize the *historical* parameter values on $(n-1)$ medical cases for a patient and evaluate their fitness on the n^{th} medical case. Finally, in Section 4.3.5 we present experimental results highlighting that use of historical parameters will lead to a higher error than using inter-patient group parameters.

4.3.1 Experimental setting

We have considered a model to predict total surface and average surface of FP for the next measurement for a patient. We have run it with the following settings - number of treatment phases is 7, however in most of cases, only 2-3 phases will contain enough information in order to run experiments.

Second main setting is the account for E2 measurements. We have launched our follicle model which takes into account E2 measurements (marked on figures by ✔ E2) and we have launched it with eliminating E2 measurements (marked on figures by ✖ E2). This experimental setting meant to verify the possibility for our model to predict total and average surface of FP, based only on previous measurement of FP and drug dose.

A structural view on experimental settings:

1. Patient-Specific Model (A)
   - model ✔ E2

     Average error on elements <total surface of FP, average surface of FP, E2>, Figure 4.1.

     Element Error on <total surface of FP>, Figure 4.2.

     Element Error on <average surface of FP>, Figure 4.3.

     Element Error on <E2>, Figure 4.4.
   - model ✖ E2

     Average error on elements <total surface of FP, average surface of FP>, Figure 4.1.

     Element Error on <total surface of FP>, Figure 4.2.

     Element Error on <average surface of FP>, Figure 4.3.
2. Inter-Patient Group Model (B.1)
   - model ✔ E2, Table 4.1a contains

     Average error on elements <total surface of FP, average surface of FP, E2>

     Element Error on <total surface of FP>

     Element Error on <average surface of FP>

     Element Error on <E2>
   - model ✖ E2, Table 4.1b contains

     Average error on elements <total surface of FP, average surface of FP>

     Element Error on <total surface of FP>

     Element Error on <average surface of FP>
3. Methodology Validation (B.2)
   - model ✔ E2, Table 4.3a contains

     Average error on elements <total surface of FP, average surface of FP, E2>

     Element Error on <total surface of FP>

     Element Error on <average surface of FP>

     Element Error on <E2>
   - model ✖ E2, Table 4.3b contains

     Average error on elements <total surface of FP, average surface of FP>

Element Error on <total surface of FP>
Element Error on <average surface of FP>

4. Comparing Historical Prediction to a Group Prediction (C)

- model ✔ E2

  Average error on elements <total surface of FP, average surface of FP, E2>
  (a) Historical Prediction, Figure 4.6.
  (b) Group Prediction, Figure 4.5.
  Element Error on <total surface of FP>
  (a) Historical Prediction, Figure 4.8.
  (b) Group Prediction, Figure 4.7.
  Element Error on <average surface of FP>
  (a) Historical Prediction, Figure 4.10.
  (b) Group Prediction, Figure 4.9.
  Element Error on <E2>
  (a) Historical Prediction, Figure 4.12.
  (b) Group Prediction, Figure 4.11.

- model ✖ E2

  Average error on elements <total surface of FP, average surface of FP>
  (a) Historical Prediction, Figure 4.6.
  (b) Group Prediction, Figure 4.5.
  Element Error on <total surface of FP>
  (a) Historical Prediction, Figure 4.8.
  (b) Group Prediction, Figure 4.7.
  Element Error on <average surface of FP>
  (a) Historical Prediction, Figure 4.10.
  (b) Group Prediction, Figure 4.9.

We discuss each of four experiments, listed above, in Sections 4.3.2, 4.3.3, 4.3.4, 4.3.5 respectively.

4.3.2 Patient-Specific Model (A)

The patient-specific model is the model predicting the follicle dynamics solely for a patient. Thus, the patient-specific error is the error obtained for a patient taking into account all available medical cases. Whereas patient-specific model is parameter dependent model, we obtain parameter values by minimizing the error between measurements and predictions, using the AMPL tool to solve quadratic optimization problems. As we chose to minimize two types of errors - average on elements <total surface of FP, average surface of FP> and error on elements <surface total of FP>, <surface average of FP>, along with choice to both account for E2 measurements and not, it led to four types of experiments.

Two types of average error are demonstrated on Fig. 4.1. One is the average on elements <total surface of FP, average surface of FP, E2> and other one is the average on elements <total surface of FP, average surface of FP>. The only distinction between them is the account for E2 measurements in the model. Model, taking into account E2 measurements, has more than 60% of patients with error less then 10%

and more then 80% of patients with error less then 20%. Model, which does not take into account E2 measurements, has more than 40% of patients with error less then 10% and more then 75% of patients with error less then 20%. This patient-specific experiment showed us that model which does not take into account E2 measurements provides prediction almost as good as model, taking into account E2 measurements, thus, we can obtain low-error prediction for total and average surface of FP, based on previous measurement of FP and drug dose of FSH/LH or FSH.

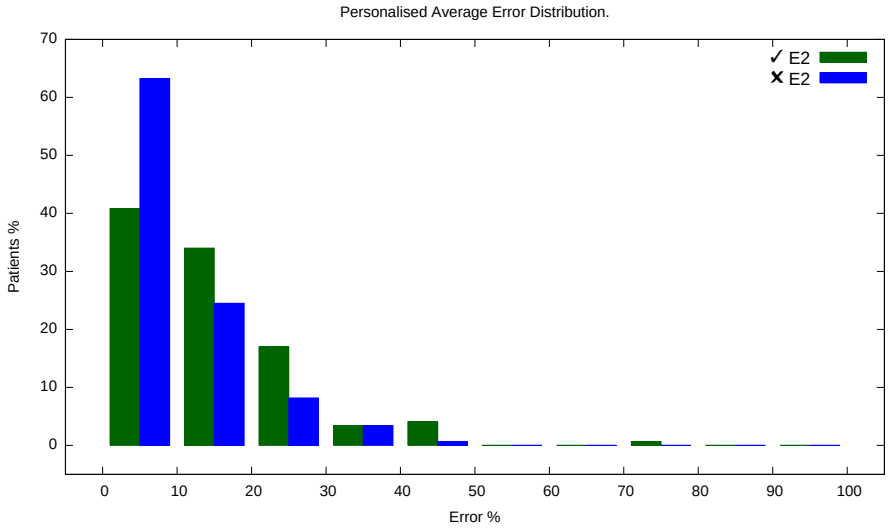

FIGURE 4.1: Shows distribution of % error, obtained for patient-specific cases (see Eq. 4.4) by optimizing an average error on elements <total surface of FP, average surface of FP, e2> (green bars) and by optimizing an average error on elements <total surface of FP, average surface of FP> (blue bars).

Second type of error, obtained seperatly for each element <total surface of FP>, <average surface of FP>, <E2> is presented on Figures 4.2, 4.3, 4.4. As discussed earlier, three seperatly obtained errors on elements <total surface of FP>, <average surface of FP>, <E2> are all calculated in two manners - taking into account E2 measurements or not. Figure 4.2 shows same summary as Figure 4.1, we can obtain low-error prediction for <total suface of FP>, based on previous measurement of FP and drug dose of FSH/LH or FSH. More than 50% of patients have error less than 10%, it applies to both versions of model. And more than 80% of patients have less then 20% error.

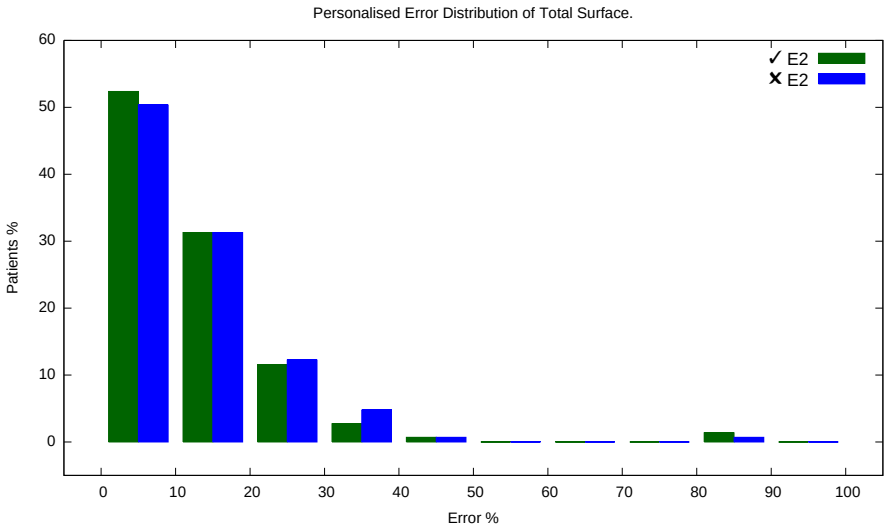

FIGURE 4.2: Shows distribution of % error, obtained for patient-specific cases (see Eq. 4.5) by optimizing error on <total surface of FP> with use of E2 measurements (green bars) and by optimizing error on <total surface of FP> without use of E2 measurements (blue bars).

Close values of patient-specific error shows <average surface of FP> element (Fig. 4.3) to error values on <total surface of FP> (Fig. 4.2).

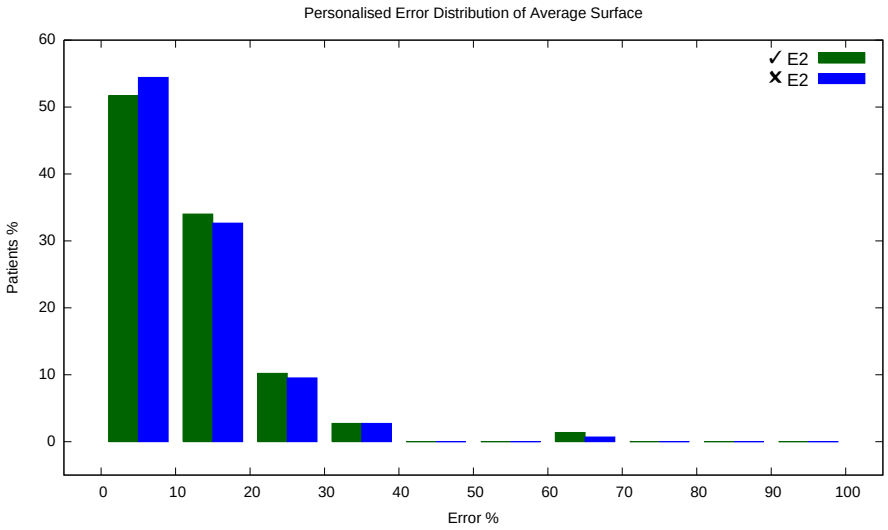

FIGURE 4.3: Shows distribution of % error, obtained for patient-specific cases (see Eq. 4.5) by optimizing error on <average surface of FP> with use of E2 measurements (green bars) and by optimizing error on <average surface of FP> without use of E2 measurements (blue bars).

Error, obtained seperatly for an <E2> element is presented on Figure 4.4. Almost 40% of patients have error less than 10% and 65% of patients have less than 20% error.

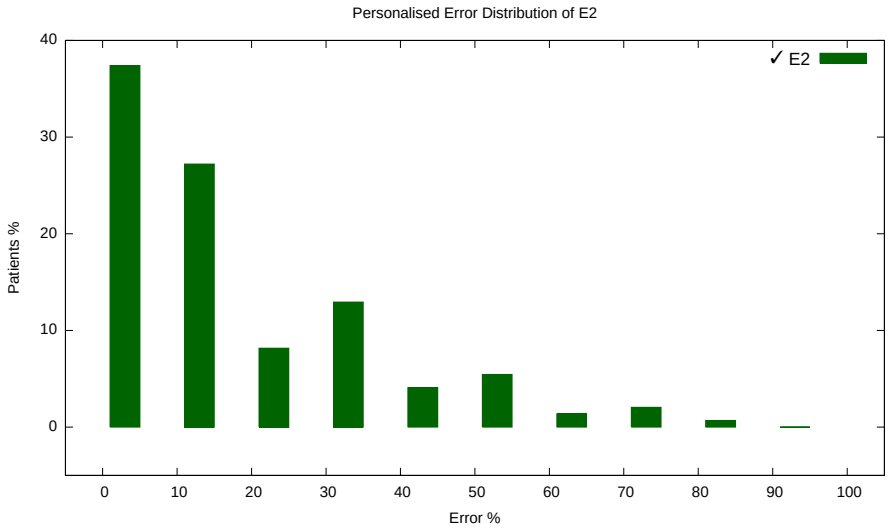

FIGURE 4.4: Shows distribution of % error, obtained for patient-specific cases (see Eq. 4.5) by optimizing error on <E2>.

The experimental results for *patient specific* error have been performed in order to validate our model with respect to a patient, including all previous patient measurements. As a result, model, taking into account E2 measurements, revealed that more then 80% of patients with error less then 20%. While model excluding E2 measurements provided more then 75% of patients with error less then 20%. Both type of models have low error values, which testifies the correctenes of our model. The experimental results for patient specific error were obtained on data from 147 patients with at least two medical cases, where each treatment phase has at least two measurements.

4.3.3 Inter-Patient Group Model (B.1)

The inter-patient group model is the model predicting the follicle dynamics solely on patients with external factors same as a patient group. We obtain inter-patient group parameters values for each group seperatly. Due to the experimental setting we get four type of errors. First minimizes an average error (see Section 4.2.4.1) on elements <surface total, surface average, E2> (I), second minimizes as well an average error on the same model, yet eliminating E2 measurements (II). Third type minimizes error by elements <surface total>, <surface average>, <E2> (III), see Section 4.2.4.2. Last type minimizes as well error by elements, yet eliminating E2 measurements (IV). In order to solve an optimization problem and obtain parameter values minimizing one of four types error, we use AMPL tool (B.1).

	Deterministic Error + **E2**			
	Optimizing Error by Element, + **E2**			Optimizing AVG Error, + **E2**
Patient Group	S.Total error (%)	S.Avg error (%)	E2 level error (%)	AVG Error (%)
1	20.0	16.8	18.0	16.1
2	13.9	11.0	16.5	13.0
3	13.6	14.0	17.4	13.4
4	12.0	12.6	11.2	10.9
5	22.5	20.9	21.4	21.2
6	19.2	18.2	22.5	21.4
7	22.2	20.0	24.3	27.9
8	19.8	15.2	24.4	18.4
9	14.9	15.2	17.2	16.9

(A) This table shows values of deterministic error (see Section 4.2.4) (B.1) for our follicle model retaining E2 measurements. It optimizes parameters for an average error on elements <surface total, surface average, E2> and error by elements <surface total>, <surface average>, <E2>, see Eq. (4.4) and (4.5) respectively. See Patient Group description in Table 4.2.

	Deterministic Error - **E2**		
	Optimizing Error by Element, - **E2**		Optimizing AVG Error, - **E2**
Patient Group ID	S.Total error (%)	S.Avg error (%)	AVG Error (%)
1	18.9	16.9	20.8
2	8.3	7.5	12.7
3	14.7	17.6	15.2
4	25.3	21.0	22.0
5	21.9	23.0	19.3
6	19.4	22.6	19.6
7	19.3	15.7	17.9
8	17.4	16.2	16.2
9	15.2	12.1	18.0

(B) This table shows values of deterministic error (see Section 4.2.4) (B.1) for our follicle model, optimizing parameters for average error and error by elements without E2 level, see Eq. (4.4) and (4.5) respectively. See Patient Group description in Table 4.2.

Patient Group ID	Patient Group
1	<MR, H, FSH/LH, Id.1>
2	<MR, H, FSH/LH, Id.2>
3	<MR, E, FSH/LH, Id.1>
4	<MR, E, FSH/LH, Id.3>
5	<ER, H, FSH/LH, Id.1>
6	<ER, H, FSH/LH, Id.2>
7	<ER, E, FSH/LH, Id.1>
8	<ER, I, FSH/LH, Id.1>
9	<HR, H, FSH/LH, Id.1>

TABLE 4.2: Contains Patient Group description and a corresponding to it Patient Group ID.

Tables 4.1a, 4.1b contain experimental results validating our model with respect to a patient group (B.1). They compare error types (I) and (III), (II) and (IV) respectively. If we compare error values on element <surface total> in model using E2 measurements (Table 4.1a) and the one eliminating them (Table 4.1b), it is cleat that only one patient group significantly wins from use of model using E2 measurements. This group has <MR, E, FSH/LH, Id.3> as external factors. All other groups has similar error values on element <surface total> in between model using E2 measurements and model eliminating them.

Group with same external factors <MR, E, FSH/LH, Id.3> also significantly wins from use of model using E2 measurements, on element <surface average>, if we compare error values on element <surface averafe> in model using E2 measurements (Table 4.1a) and the one eliminating them (Table 4.1b).

The experimental results for model optimizing *inter-patient group* parameters clearly shows that our model is capable to predict total and average surface of follicle profile based exclusively on preceding measurement of FP and stimulation drug dose.

4.3.4 Methodology Validation (B.2)

In this Section we validate our methodology with respect to a patient group (B.2), by applying leave-one-out bootstrap technique (see Section 2.2.5, 4.2.5) and obtaining bootstrap average error and standard deviation. As in Section 4.3.3 we obtain four types of error, average by elements (error types I, II) and error on elements (error types III, IV). Error types I and II obtained with respect to Eq. 4.9, 4.10, while error types III and IV obtained with respect to Eq. 4.7, 4.8.

Table 4.3a contains error values corresponding to the error types I and III. At minimum error type III for <surface total> element has value of 21% in group with external factors <MR, H, FSH/LH, Id.2> and at maximum 37.2% in <HR, H, FSH/LH, Id.1> group. Element <surface average> has the lowest error in the same group as <surface total> and maximum at 30.1% in <ER, E, FSH/LH, Id.1> group. Element <E2> has higher errors in comparison to <surface total> and <surface average> elements. At minimum it has 23.4% in <MR, H, FSH/LH, Id.1> group and at maximum it has 45.4% in <ER, I, FSH/LH, Id.1> group, both of which are higher than in <surface total> and <surface average>. Standard deviation various through patient groups. On average it is 27% on <surface total> element and 24% for <surface average> element. The highest, averaged through groups, standard deviation is 28% on <E2> element. Error type III is average error on elements <surface total, surface

average, E2>, it has at minimum value of 21.3% in <MR, H, FSH/LH, Id.1> group, while at maximum 34.3% in <ER, E, FSH/LH, Id.1> group.

Table 4.3b contains error values corresponding to the error types II and IV. If we compare Table 4.3a to Table 4.3b, we can see that in general validation for model eliminating E2 measurements provides lower errors. For example, error type I on element <surface total> is 21% at minimum in patient group <MR, H, FSH/LH, Id.2> (Table 4.3a), while error type II (Table 4.3b) at minimum is on the same group and equals to 20.3%. On average through patient groups error type II on element <surface total> is 25% and on <surface average> element it is around 21%.

Although, validation of model taking into account E2 measurements gives relatively low error values, the methodoly validation results show higher values error in comparison to model eliminating E2 measurements. The methodoly validation experimental results clearly shows that our model performs best to predict total and average surface of follicle profile based exclusively on preceding measurement of FP and stimulation drug dose. Yet, one group with external factors <MR, E, FSH/LH, Id.3> still has as unpredictable behaviour due to lack of precise measurements.

	Bootstrap Error + **E2**			
	Optimizing Error by Element, + **E2**			Optimizing AVG Error, + **E2**
Patient Group ID	S.Total error (%) / S.Total St.D (%)	S.Avg error (%) / S.Avg St.D (%)	E2 level error (%) / E2 level St.D (%)	AVG Error (%) / AVG St.D (%)
1	23.3/14	21.6/16	23.4/15	21.3/23
2	21.0/23	20.0/25	36.1/35	24.3/33
3	24.5/21	29.4/21	33.4/24	32.2/38
4	30.8/38	29.1/34	25.5/40	32.7/39
5	30.0/28	27.2/48	26.5/22	25.8/29
6	27.3/56	28.1/17	33.3/26	30.1/33
7	30.0/19	30.1/17	35.2/23	34.3/37
8	36.2/22	26.7/18	45.4/46	32.4/35
9	37.2/32	28.9/23	34.5/27	33.1/35

(A) This table shows values of bootstrap error and standard deviation for the follicle model (see Section 4.2.3). It optimizes parameters for an average error on elements <surface total, surface average, E2> and error by elements <surface total>, <surface average>, <E2>, see Eq.(4.9, 4.10) and (4.7, 4.8) respectively. See Patient Group description in Table 4.2.

	Bootstrap Error - **E2**		
	Optimizing Error by Element, - **E2**		Optimizing AVG Error, - **E2**
Patient Group ID	S.Total error (%) / S.Total St.D (%)	S.Avg error (%) / S.Avg St.D (%)	AVG Error (%) / AVG St.D (%)
1	23.4/13	21.2/13	23.1/24
2	20.3/19	16.9/21	22.0/30
3	28.9/21	28.5/ 19	29.8/31
4	50/47	38.1/26	44.6/46
5	28.9/26	23.4/16	24.8/27
6	27.8/17	28.5/17	28.5/30
7	34.2/19	29.1/20	30.4/33
8	33.2/20	23.9/14	30.8/34
9	36.9/27	29.9/26	33.4/38

(B) This table shows values of bootstrap error and standard deviation for the follicle model (see Section 4.2.3). It optimizes parameters for an average error on elements <surface total, surface average> and error by elements <surface total>, <surface average>, see Eq.(4.9, 4.10) and (4.7, 4.8) respectively. See Patient Group description in Table 4.2.

4.3.5 Comparing Historical Prediction to a Group Prediction (C)

In this Section we investigate if with the use of a previous meadical cases of a patient, we can predict patient behaviour on a future medical case (C). In order to do so, we obtain all medical cases available of a patient and by minimizing the difference between measurements of elements and estimations as an error, we obtain parameter values, let us call them *historical* parameters.

In order to compare historical prediction to the inter-patient group prediction (C) we first use *historical* parameters to calculate four types of errors on a last available medical case for each patient (last available medical case was not used in optimization problem to obtain *historical* parameters). Second use *inter-patient group* parameters (see Section 2.2.4) to calculate four types of errors on a last available medical case for each patient.

First type of error minimizes an average error (see Section 4.2.4.1) on elements <surface total, surface average, E2> (I), second minimizes as well an average error on the same model, yet eliminating E2 measurements (II). Third type minimizes error by elements <surface total>, <surface average>, <E2> (III), see Section 4.2.4.2. Last type minimizes as well error by elements, yet eliminating E2 measurements (IV).

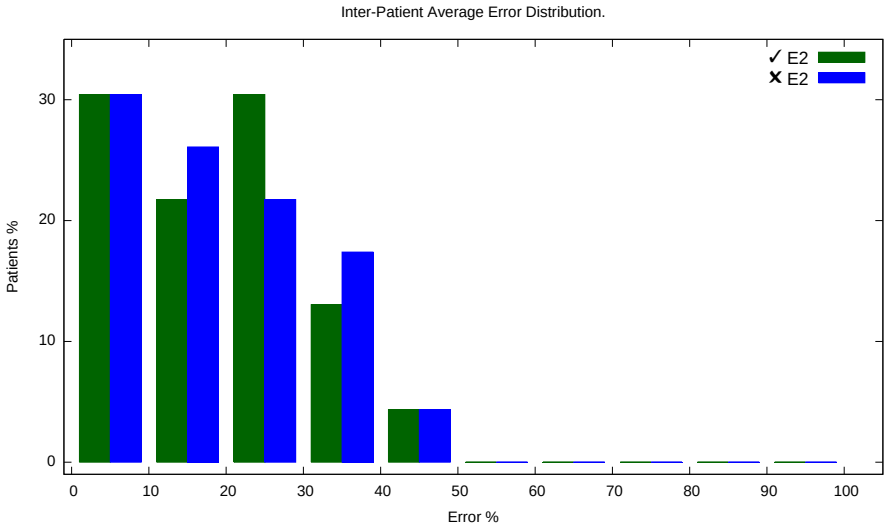

FIGURE 4.5: Shows distribution of % error, obtained by minimizing an average error on elements <total surface of FP, average surface of FP, e2> (green bars) and by optimizing an average error on elements <total surface of FP, average surface of FP> (blue bars). Predictions were obtained based on inter-patient group parameters for historical cases (see Eq. 4.4) solely on the last medical case of each patient.

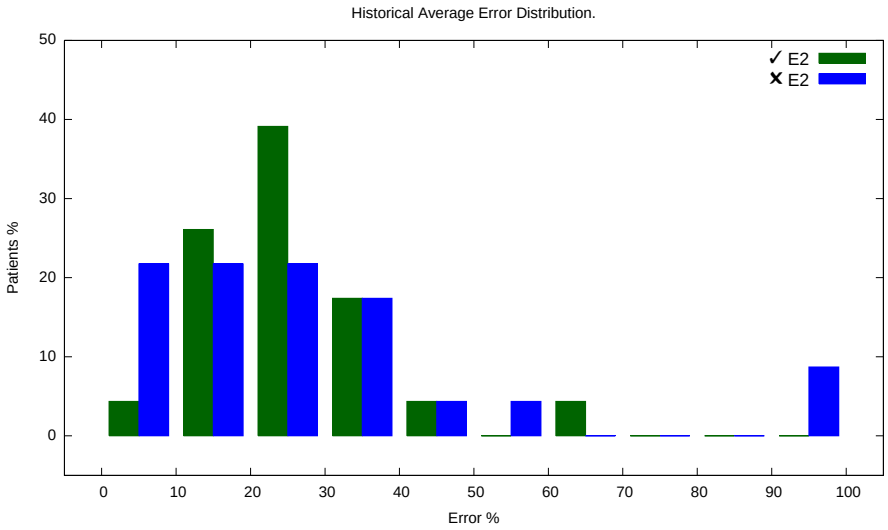

FIGURE 4.6: Shows distribution of % error, for patients with multiple medical cases (see Eq. 4.4), obtained by optimizing an average error on elements <total surface of FP, average surface of FP, e2> (green bars) and by optimizing an average error on elements <total surface of FP, average surface of FP> (blue bars).

Figures 4.5 and 4.6 shows two types of error I and II. The difference between them is that Figure 4.6 shows distribution of error calculated using historical parameters, while Figure 4.5 shows distribution of error calculated using inter-patient group parameters. The resemblance between them is that both type of parameters were used to calculated the prediction for the last available medical case of a patient. If one compares Figures 4.5 and 4.6, he or she will see that distribution using inter-patient group parameters has almoust 55% of historical patients with error value less than 20%, while distribution using historical parameters has less than 30% of historical patients with error value less than 20% (model using E2 measurements, shown with green bars at the graphs). Same tendency shows model eliminating E2 measurements (shown with blue bars at the graphs).

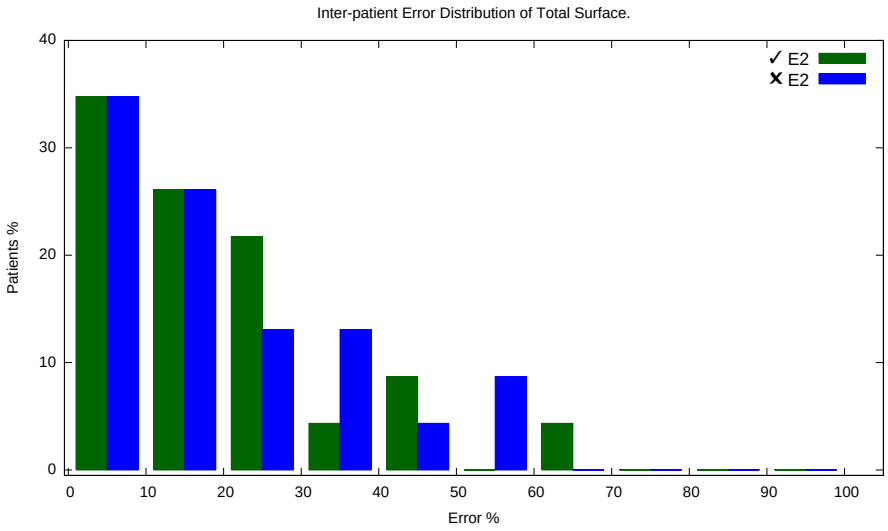

FIGURE 4.7: Shows distribution of % error, obtained by minimizing error on element <total surface of FP> with use of E2 measurements (green bars) and by optimizing a error on element <total surface of FP> eliminating E2 measurements (blue bars). Predictions were obtained based on inter-patient group parameters for historical cases (see Eq. 4.5) solely on the last medical case of each patient.

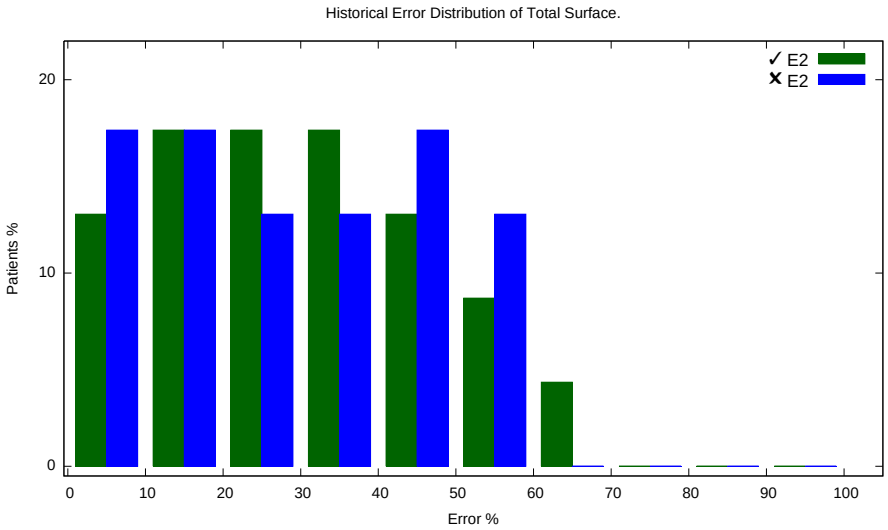

FIGURE 4.8: Shows distribution of % error, for patients with multiple medical cases (see Eq. 4.5), obtained by optimizing error on <total surface of FP> with use of E2 measurements (green bars) and by optimizing error on <total surface of FP> without use of E2 measurements (blue bars).

Figures 4.7 and 4.8 shows two types of error on element <surface total>, only now for type errors III and IV. If we compare those figures we will see that distribution using inter-patient group parameters has more than 60% of patients with error less than 20%, while distribution using historical parameters less than 30% of patients with the same 20% error. Moreover, we should highlight that error distribution using historical parameters (Figure 4.8) has a much larger bell. If we take a look at error in between 50% - 60%, we will see less than 10% of patients in model using E2 measurements has it and model eliminating E2 has more than 10% of patients in it. However, if we take a look at distribution using inter-patient group parameters (Figure 4.7) it has 0% and less than 10% of patients for a model using E2 measurements and for a one eliminating those.

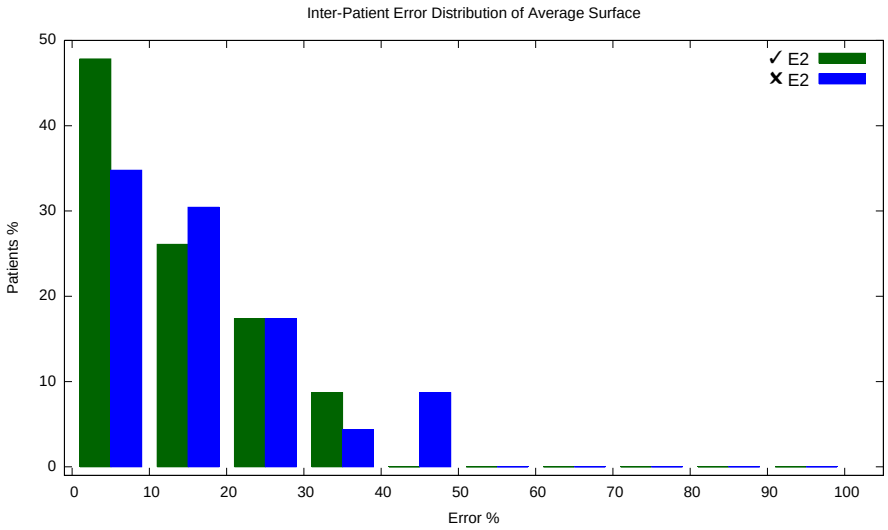

FIGURE 4.9: Shows distribution of % error, obtained by minimizing error on element <average surface of FP> with use of E2 measurements (green bars) and by optimizing a error on element <average surface of FP> eliminating E2 measurements (blue bars). Predictions were obtained based on inter-patient group parameters for historical cases (see Eq. 4.5) solely on the last medical case of each patient.

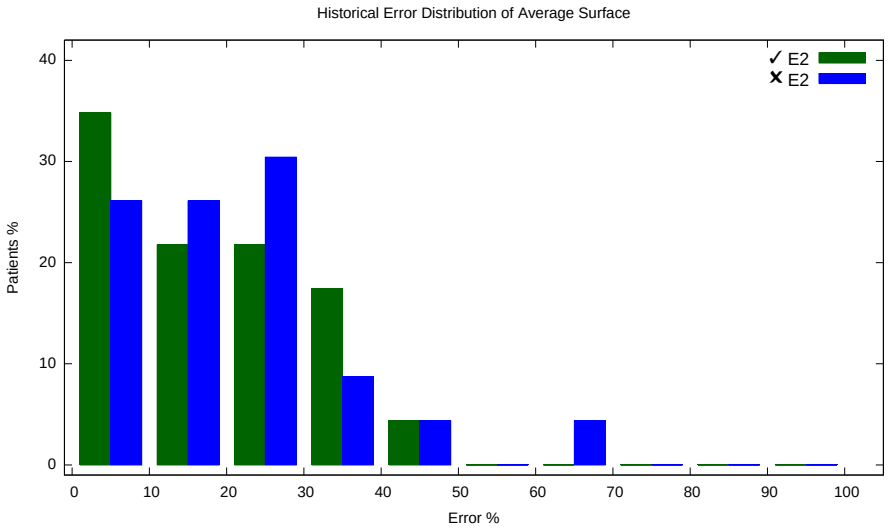

FIGURE 4.10: Shows distribution of % error, for patients with multiple medical cases (see Eq. 4.5), obtained by optimizing error on <average surface of FP> with use of E2 measurements (green bars) and by optimizing error on <average surface of FP> without use of E2 measurements (blue bars).

Figures 4.9 and 4.10 shows same two types of error III and IV, only now on element <surface average>. If we compare those figures we will see that distribution (on model taking into consideration E2 measurements) using inter-patient group parameters has more than 70% of patients with error less than 20%, while distribution using historical parameters less than 55% of patients with the same 20% error. Same tendency has model eliminating E2 measurements. Distribution using inter-patient group parameters has more than 65% of patients with error less than 20%, while distribution using historical parameters less than 50% of patients with the same 20%

error. Also, we should highlight that error distribution, obtained from historical parameters, on element <surface average> has a less wide bell in comparison to the distribution on <surface total> element. If we take a look at error in between 50% - 60%, we will see that both models, as well as both distributions (inter-patient group and historical) has 0% of patient in this range.

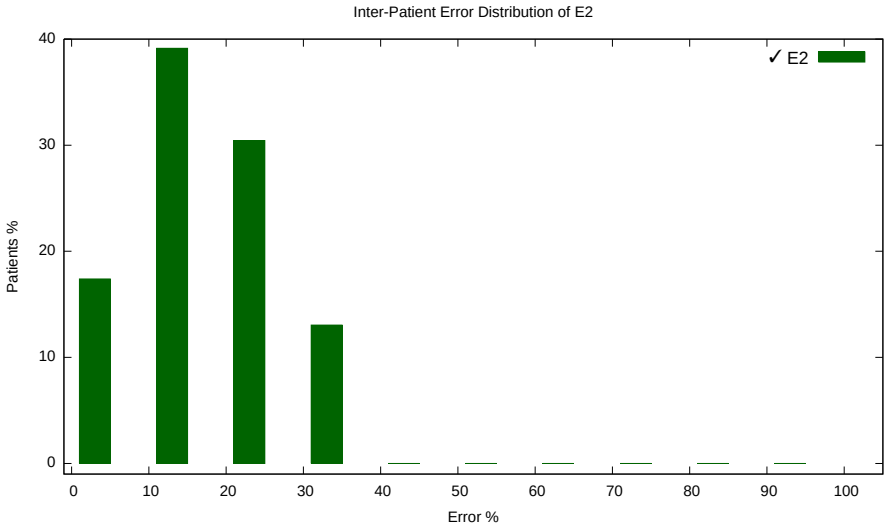

FIGURE 4.11: Shows distribution of % error, obtained by minimizing error on element <E2> (green bars). Predictions were obtained based on inter-patient group parameters for historical cases (see Eq. 4.5) solely on the last medical case of each patient.

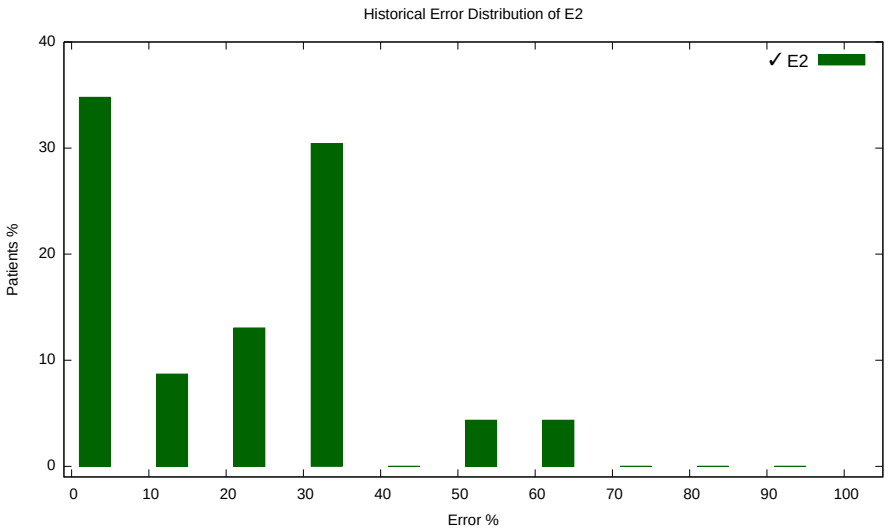

FIGURE 4.12: Shows distribution of % error, for patients with multiple medical cases (see Eq. 4.5), obtained by optimizing error on <E2>.

Figures 4.11 and 4.12 shows same two types of error III and IV, only now on element <E2>. If we compare those figures we will see that distribution using inter-patient group parameters has around 60% of patients with error less than 20%, while distribution using historical parameters less than 45% of patients with the same 20% error. Moreover, if we take a look at error in between 50% - 60%, we will see that distribution basen on inter-patient group parameters has 0% of patients in this range, while distribution basen on historical parameters has around 5% of patients in it.

This Section allowed us to investigate if using previous meadical cases of a patient except a current one, and obtaining *historical* parameter values, could be used to predict patient behaviour on a current medical cases. As experiments showed patients behaviour changes from one medical case to another. Thus, claim for historical parameters is not viable and one should use inter-patient group parameters instead.

Chapter 5

Discussion

E2 estimations as well as follicle dynamic model are based on inter-patient group parameters. If a patient belonging to a given group $G_{<r,h,d,t>}$ behaves significantly different from what expected for patient of group $G_{<r,h,d,t>}$, then probably something is going wrong and this can suggest that further investigations are required. Secondly, during stimulation treatment, clinicians perform multiple measurements, of E2 and/or TV-US. With the help of our E2 model and follicle dynamic model clinician may reduce number both of blood samples and TV-US taken from a patient, which benefits both patient and clinian. Patients will benefit for two reasons, first she will feel more comfortable and second price of treatment will be lower. As for a clinician, it means no time delay in wayting for the results of E2 concentration from overwhelmed labs.

It is worth pointing out that, precision during TV-US is crucial for identification model parameters. Based on our experiments, measurements for groups where 2-PL outperform 1-PL, were performed by either person with Id.1 or with Id.2. This means that in those groups measurements were precise enough for our models, to detect the influence of follicle maturation stage. At the same time, measurements for the groups having an unpredictable behaviour, were also performed by person with Id.2 and Id.3. In order to have more reliable estimation models in the future, we suggest to conduct experiments based on carefully performed TV-US measurements, preferably with the use of the three-dimensional ultrasound imaging (Raine-Fenning et al., 2008).

It is only natural to ask how our results could be used in clinical practice. First of all, while it is well known that TV-US measurements are the key part in fertility treatment, while some clinicians argues whether E2 is important to measure during all stimulatment treatment or not. Most of leading clinicians agree that E2 is a sufficient part of treatment. Models for both E2 and follicles could be used as an OHSS check, but also could act as a indicator whether patient reacts to treatment as expected. Having both models combined together, provides a clinician with a full measurement prediction. Clinician who has at hand only a FP measurement, could predict a future dynamics of it using our follicle model and after obtain an E2 estimation. Second advantage of our follicle model could be an integration into TDSS and an improvement of it by predicting follicle dynamics under influence of stimulation treatment.

One application of our work, is to see it as a software package, which is given treatments and external factors, recalculates optimal parameters for multiple models. Of course, identified parameters are only as good as precision of treatment data provided to it. It is highly important to be precise while performing TV-US, in order to identify parameters. Parameter estimation could be population dependent and treatment dependent. While our data set contains treatment data gathered exclusively by Zurich Hospital, where same set of external factors is being measured

for patients, many other hospitals and clinics may follow different protocols and consider different external factors. All of it influence treatment data, thus influence parameter models. In this way, it is useful to have our package rerun the parameter identification procedure to a new given treatment set.

Another application, is to see prediction models as a calculator, which given an external factors and follicle measurement, will provide follicle dynamics and/or an esimation of E2. Let us say, a patient is being treated with FSH/LH drugs (d), healthy (h), has medium response AFC (r) and follicles are being measured by person (t), then a clinician can use our follicle model and/or one of the E2 estimation models, and can make an estimation for a given patient on a specific measurement.

Chapter 6

Conclusion

As a result of our study we have developed two families of models for E2 level estimation and a follicle dynamics model during fertiliy treatment. Table 6.1 contains an analysis of our models in terms of threats, open opportunities, weaknesses and strengths. Furthermore, the software containing the models is a part of contribution for the thesis and will be available online.

Each of our prediction models provide an estimation with an acceptable error. Which means that during fertility treatment fewer blood samples would be needed, as presently they are taken each one or two days (depending on protocol and patient responde to the treatment). Another valuable change is that at the same time estimation models would reduce treatment costs, which is another benefit for a patient. Although group parameters do not lead to accurate patient specific estimations, they are still valiable as a safety check for OHSS. In fact, a patient under treatment whose measurement levels are too different from the expected behaviour for that patient group, can be a symptom that something wrong is happening during the treatment.

Last but not least, Estradiol estimation opens up an opportunity for clinician and a patient to follow the treatment online, by using small devices which are available on the market (Fertihome). It allows patients to take Transvaginal Ultrasound by themselves at home and transmit results to clinician via Internet. Clinician, after getting measurements, may use our model to estimate Estradiol blood concentration and this way he/she will has all information needed to make a decision about next dose and/or next appointment. While E2 estimation model opens up an opportunity for healthcare at a distance, integration of it with follicle dynamics model opens up even a brighter prospective. Clinician who has at hand only a TV-US measurement, could predict a future dynamics of it using our follicle model and after obtain an E2 estimation.

Our study confirms the linear relationship between E2 concentration and surface of granulosa layer, but in a vision to have a more relaible estimation tool one should perform a conductive study, preferably using three-dimensional ultrasound imaging (Raine-Fenning et al., 2008). This approach will provide far more precise follicle measurements, thereby our models will give more precise estimations.

Model	Strength	Weakness	Opportunities	Threats
NPL	based on biological knowledge	values of the validation error are appear to be high, however are fully justified in terms of the intrinsic uncertainties in the input (see Section 2.5).	opens up an opportunity to follow the treatment online (Fertihome).	model is constructed under hypothesis that TV-US is easier to obtain with respect to E2 consentration, yet in some cases it could be the opposite.
SW	does not assume any spesific relationship between E2 and a follicle.	prediction ability is limited by the measurements precision (see Section 3.2.2.3); validation error is higher with respect to NPL family of models due to overfitting.		
Foll. Model	predict follicle dynamics based only on previous measurement and drug dose.	does not reproduce dynamics of one selected follicle.	could be integrated into TDSS.	in two patient groups fail to capture patients behaviour.

TABLE 6.1: Summary on three prediction models. Strength, Weakness, Opportunities and Threats are shown for piece-wise linear family of models, step-wise family and for follicle dynamics model.

Appendix A

Graphics from E2 hormone concentration

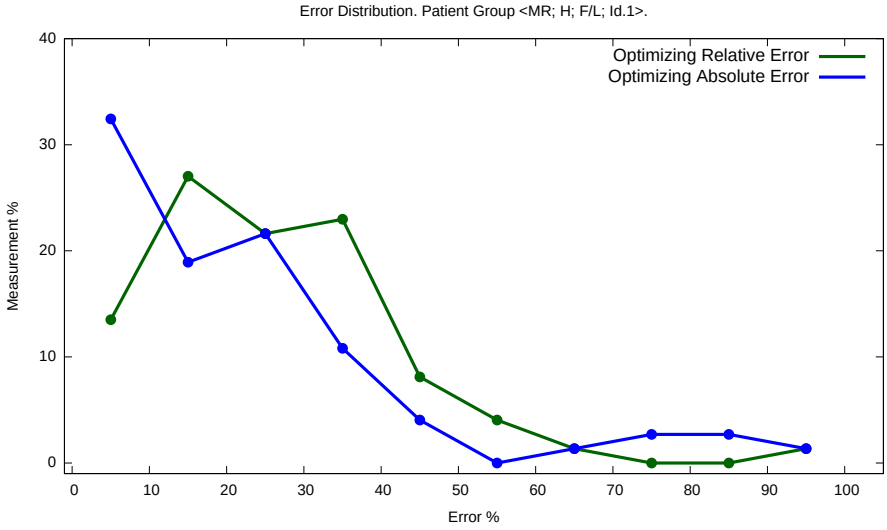

FIGURE A.1: Distribution of Deterministic Error for the patient group <MR, H, F/L, Id.1>.

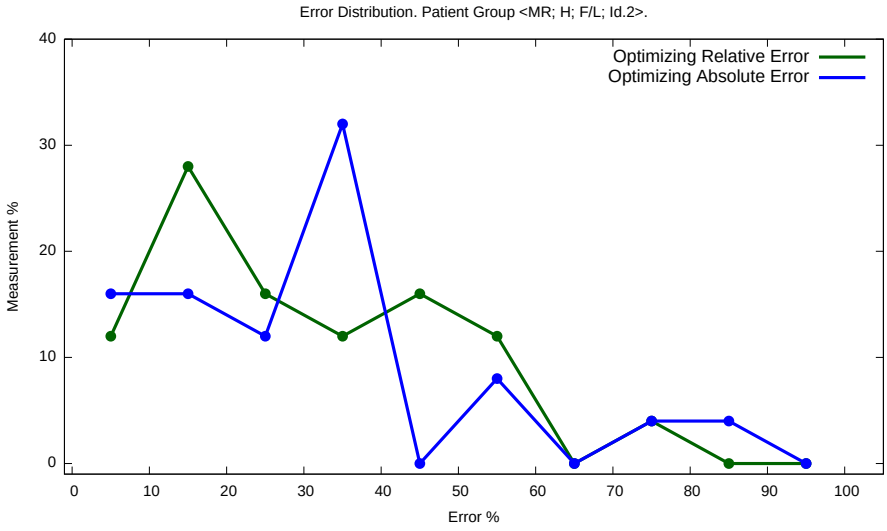

FIGURE A.2: Distribution of Deterministic Error for the patient group <MR, H, F/L, Id.2>.

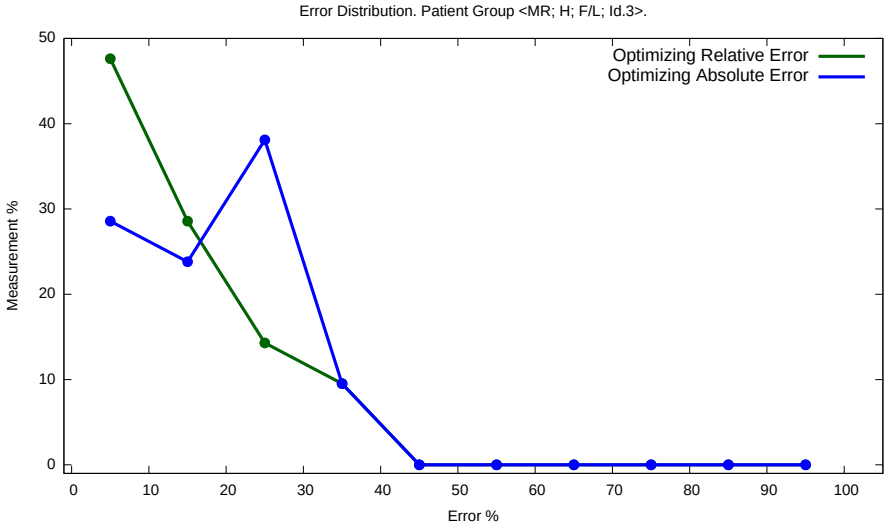

FIGURE A.3: Distribution of Deterministic Error for the patient group <MR, H, F/L, Id.3>.

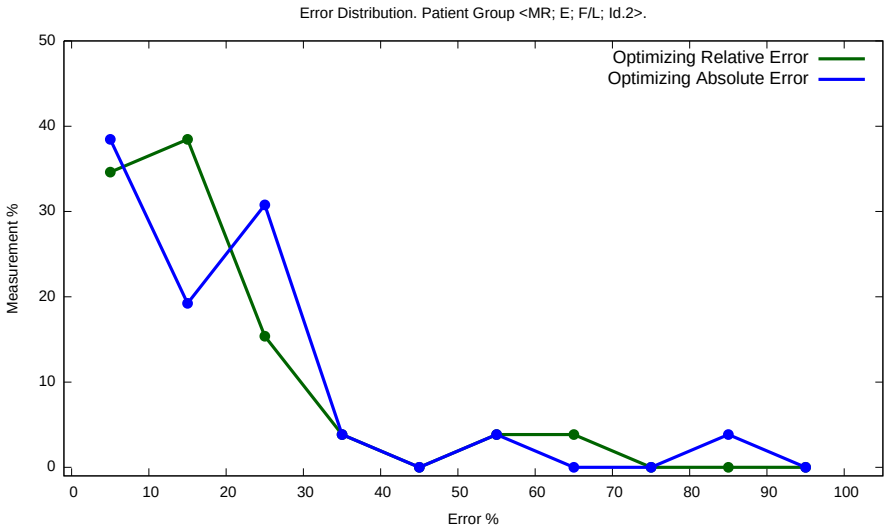

FIGURE A.4: Distribution of Deterministic Error for the patient group <MR, E, F/L, Id.2>.

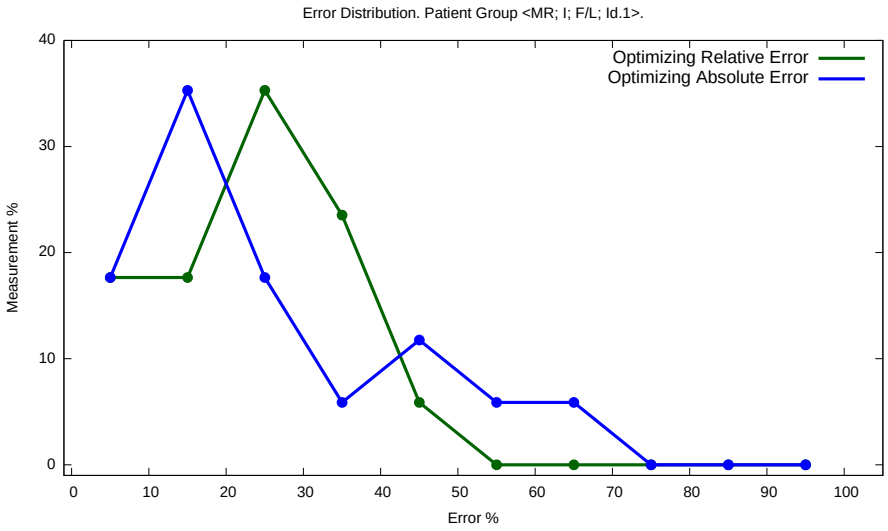

FIGURE A.5: Distribution of Deterministic Error for the patient group <MR, I, F/L, Id.1>.

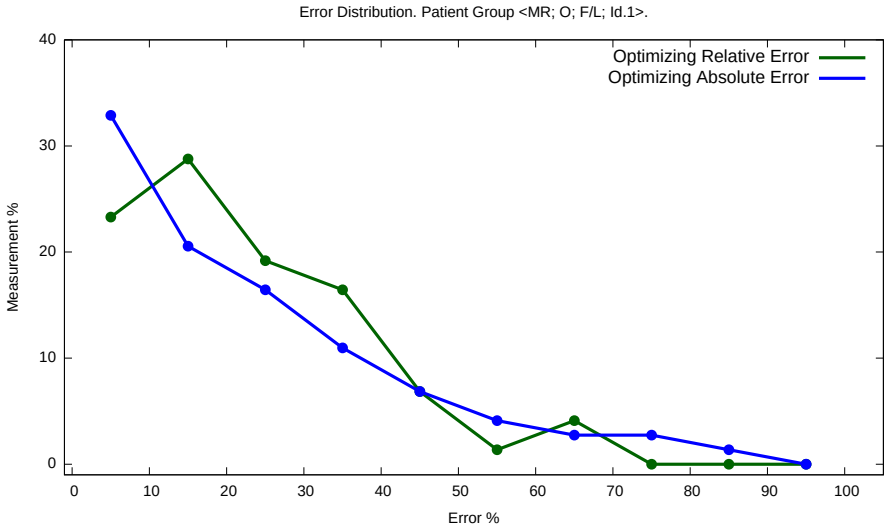

FIGURE A.6: Distribution of Deterministic Error for the patient group <MR, O, F/L, Id.1>.

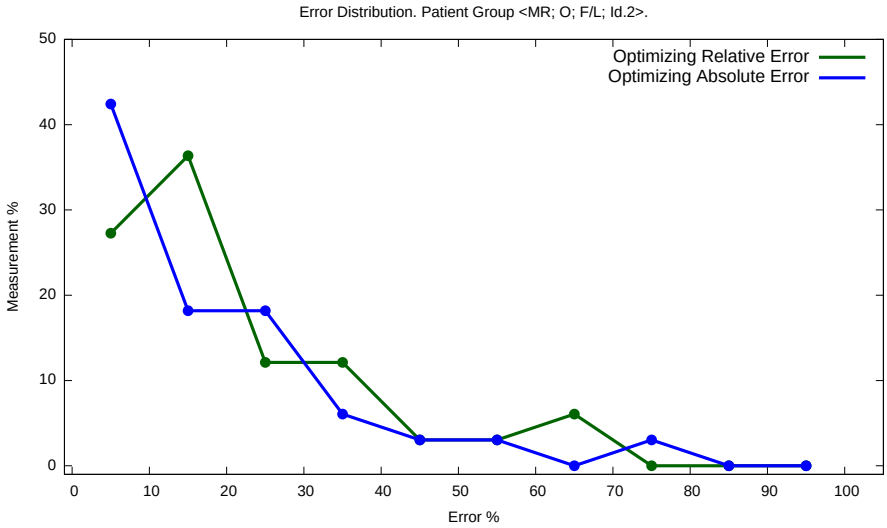

FIGURE A.7: Distribution of Deterministic Error for the patient group <MR, O, F/L, Id.2>.

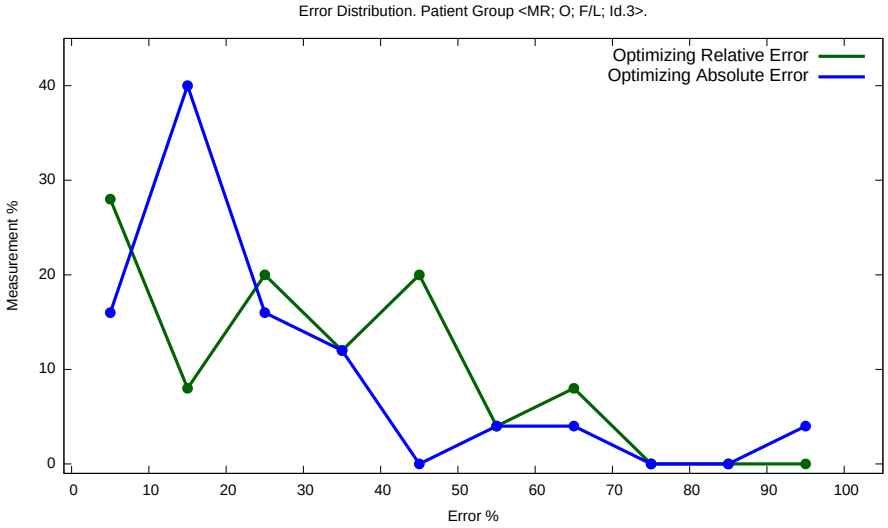

FIGURE A.8: Distribution of Deterministic Error for the patient group <MR, O, F/L, Id.3>.

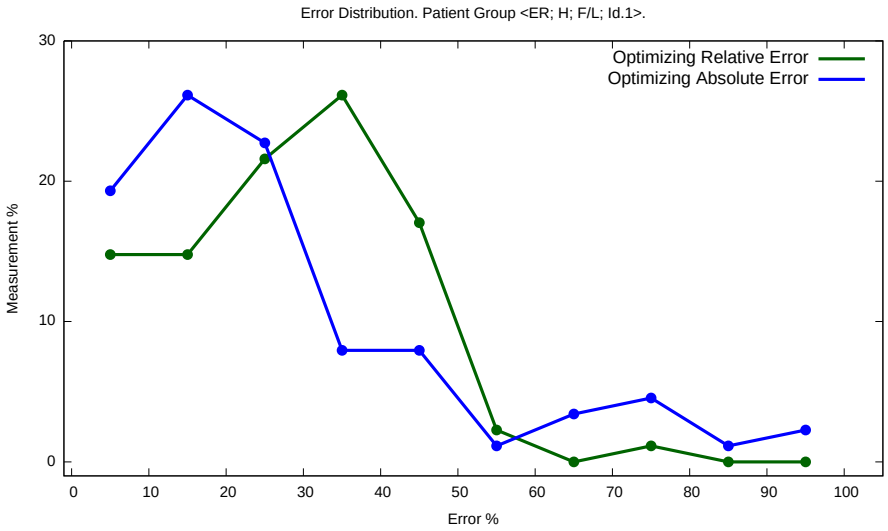

FIGURE A.9: Distribution of Deterministic Error for the patient group <ER, H, F/L, Id.1>.

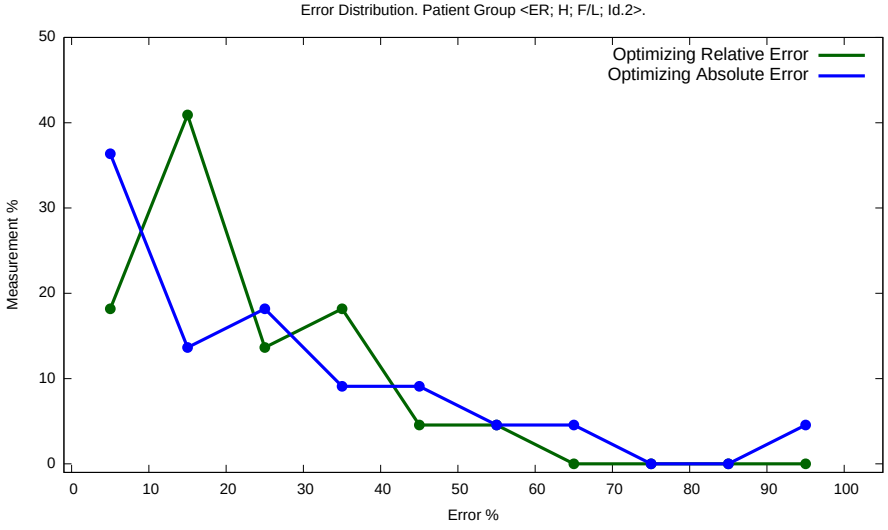

FIGURE A.10: Distribution of Deterministic Error for the patient group <ER, H, F/L, Id.2>.

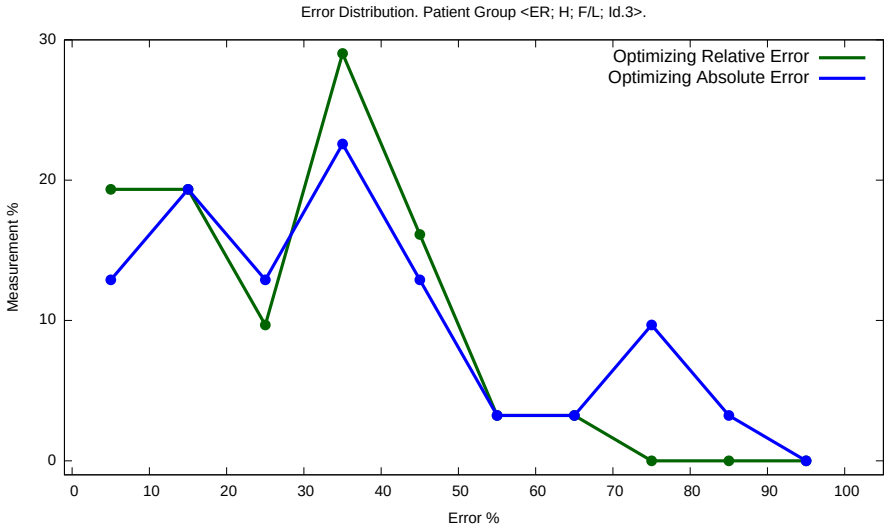

FIGURE A.11: Distribution of Deterministic Error for the patient group <ER, H, F/L, Id.3>.

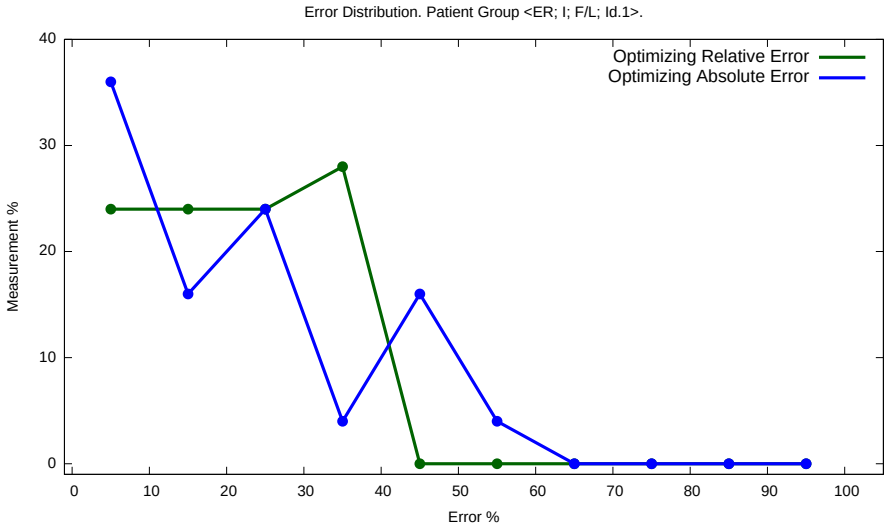

FIGURE A.12: Distribution of Deterministic Error for the patient group <ER, I, FL, Id.1>.

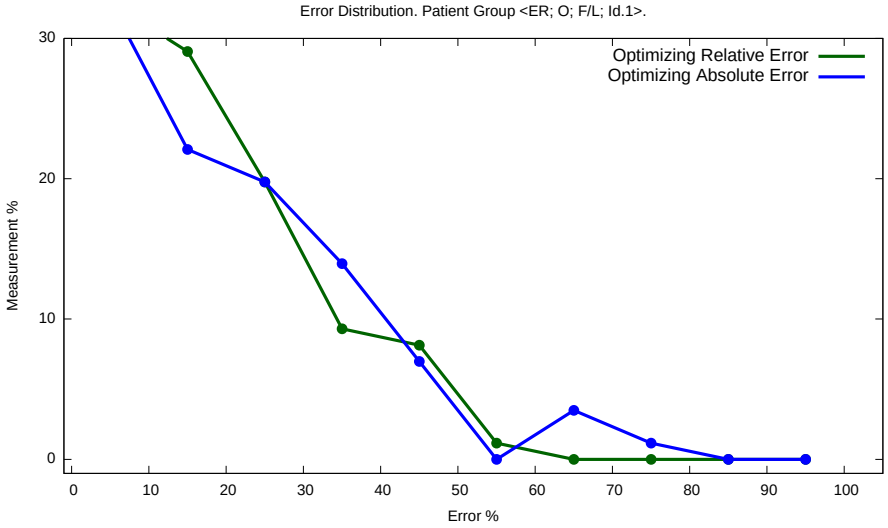

FIGURE A.13: Distribution of Deterministic Error for the patient group <ER, O, F/L, Id.1>.

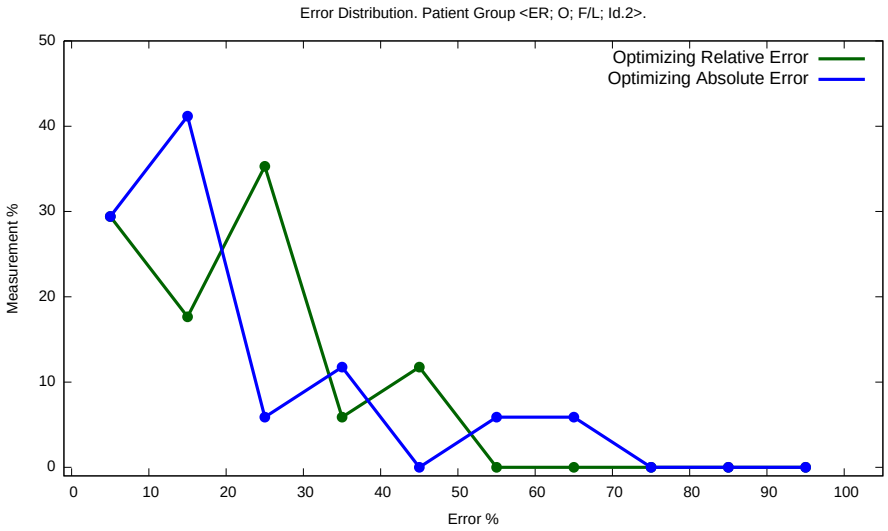

FIGURE A.14: Distribution of Deterministic Error for the patient group <ER, O, F/L, Id.2>.

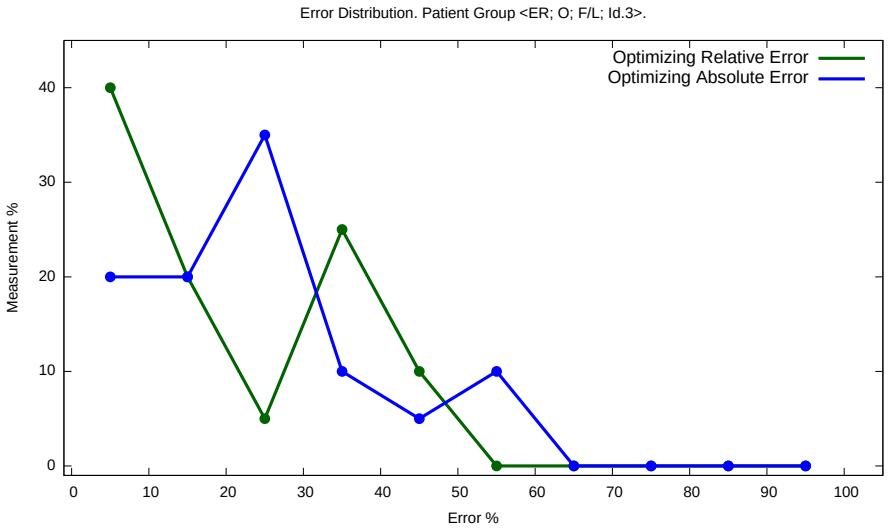

FIGURE A.15: Distribution of Deterministic Error for the patient group <ER, O, F/L, Id.3>.

Appendix B

Graphics from E2 hormone concentration Results Section: Estimation of E2 obtained by $N - PL$ family of models

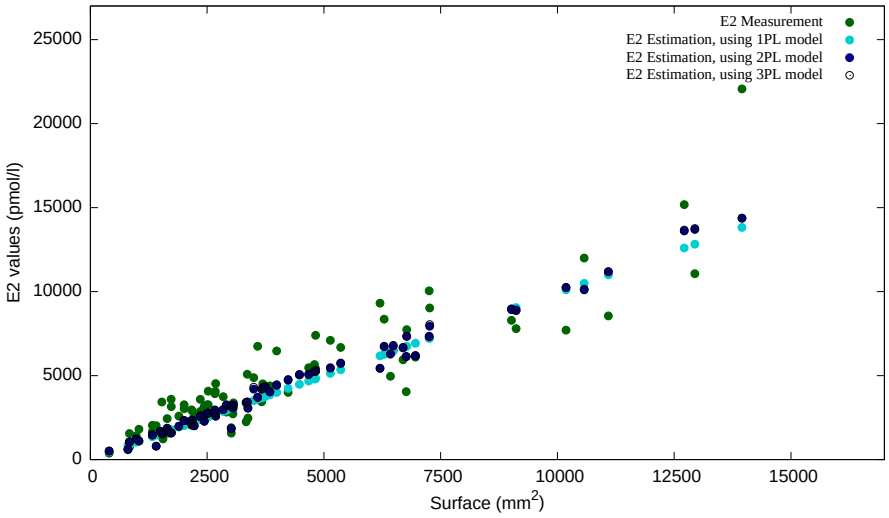

FIGURE B.1: Estimation of E2, for <MR, Healthy, FSH/LH, Id.1> group, using 1-PL, 2-PL, and 3-PL models.

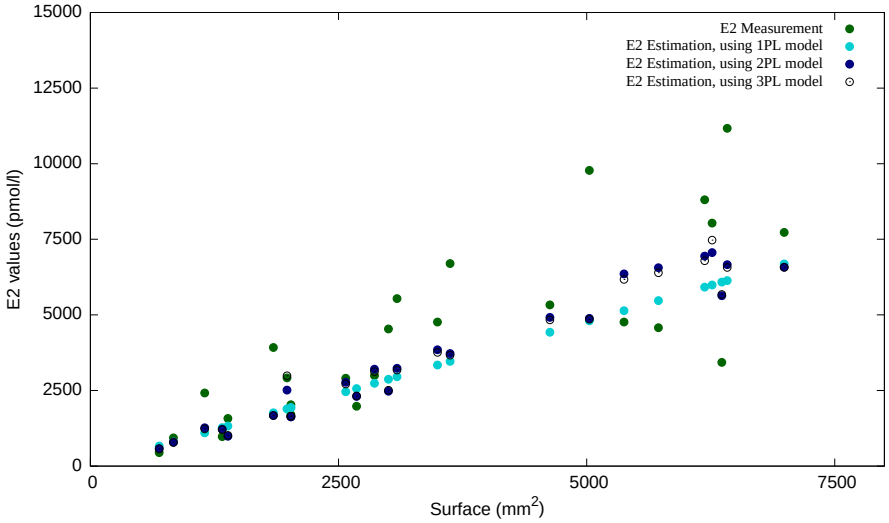

FIGURE B.2: Estimation of E2, for <MR, Healthy, FSH/LH, Id.2> group, using 1-PL, 2-PL, and 3-PL models. Group <MR, Healthy, FSH/LH, Id.2> is a group where 2-PL and 3-PL do not significantly outperform 1-PL. Measurements taken from this group are coloured with green.

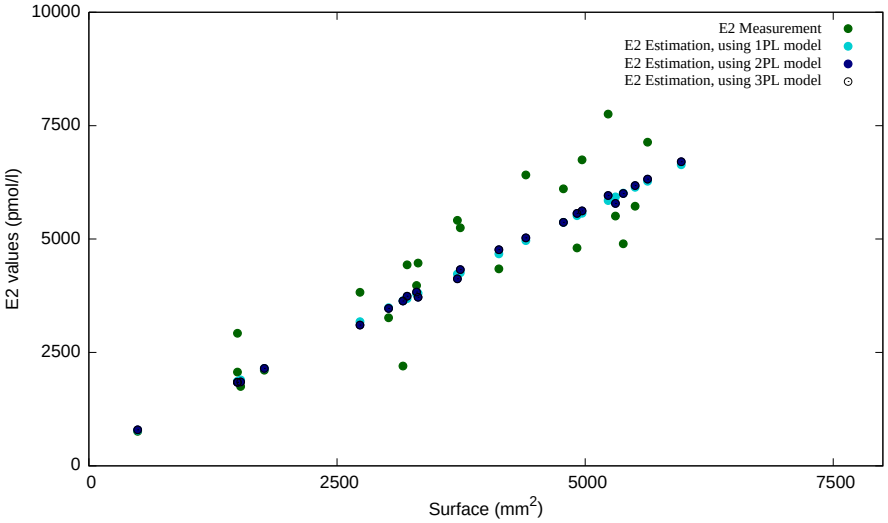

FIGURE B.3: Estimation of E2, for <MR, Endometriosis, FSH/LH, Id.1> group, using 1-PL, 2-PL, and 3-PL models. Group <MR, Endometriosis, FSH/LH, Id.1> is on the three groups, where 2-PL model outperform 1-PL. Measurements taken from this group are coloured with green.

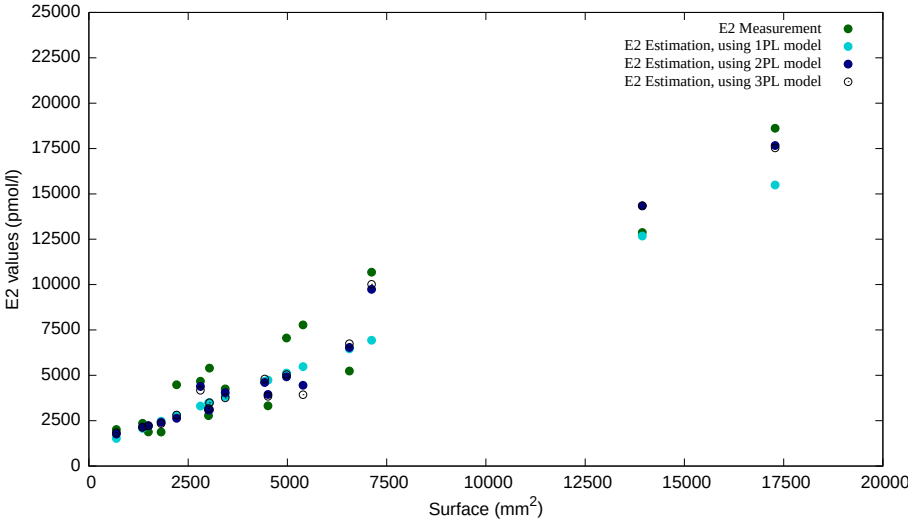

FIGURE B.4: Estimation of E2, for <MR, Idiopathic, FSH/LH, Id.1> group, using 1-PL, 2-PL, and 3-PL models.

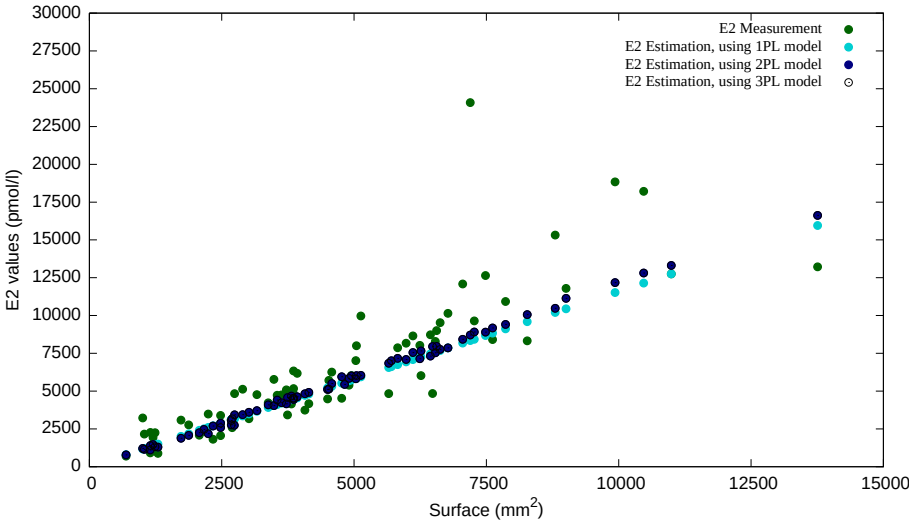

FIGURE B.5: Estimation of E2, for <MR, Other, FSH/LH, Id.1> group, using 1-PL, 2-PL, and 3-PL models.

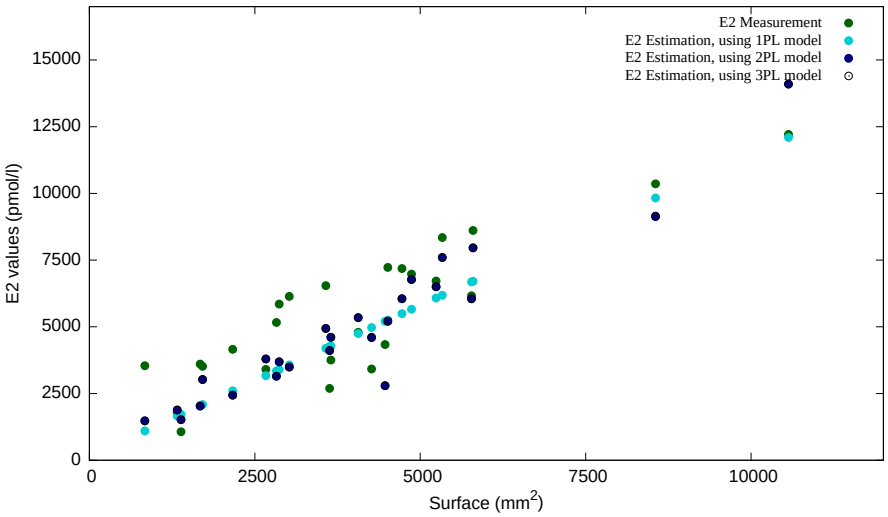

FIGURE B.6: Estimation of E2, for <MR, Other, FSH/LH, Id.3> group, using 1-PL, 2-PL, and 3-PL models.

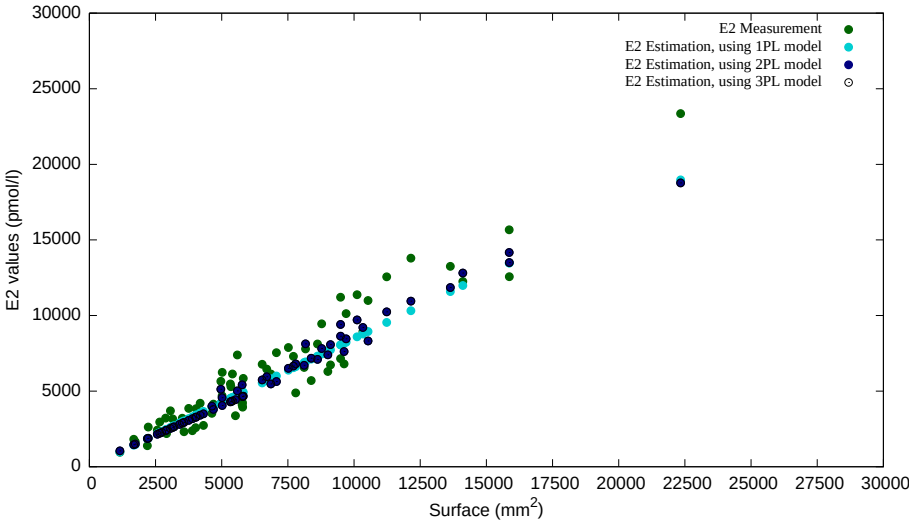

FIGURE B.7: Estimation of E2, for <ER, Healthy, FSH/LH, Id.1> group, using 1-PL, 2-PL, and 3-PL models.

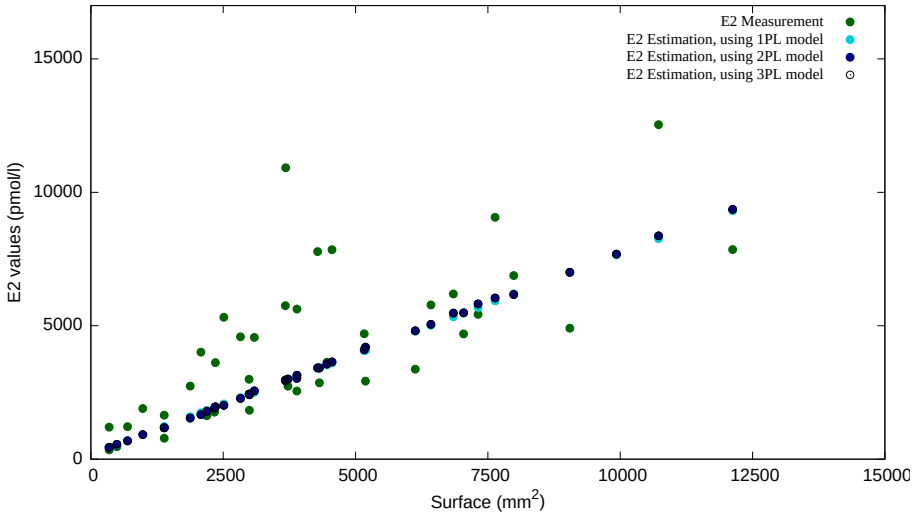

FIGURE B.8: Estimation of E2, for <ER, Healthy, FSH/LH, Id.3> group, using 1-PL, 2-PL, and 3-PL models.

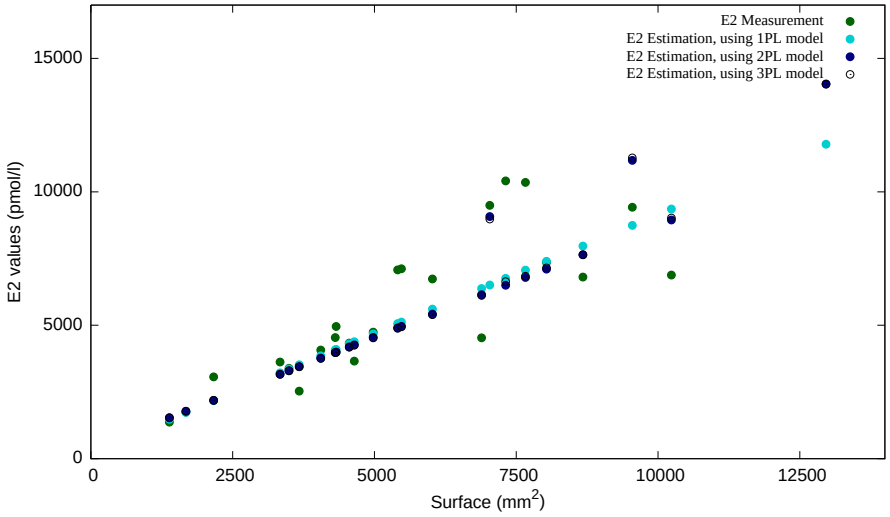

FIGURE B.9: Estimation of E2, for <ER, Idiopathic, FSH/LH, Id.1> group, using 1-PL, 2-PL, and 3-PL models.

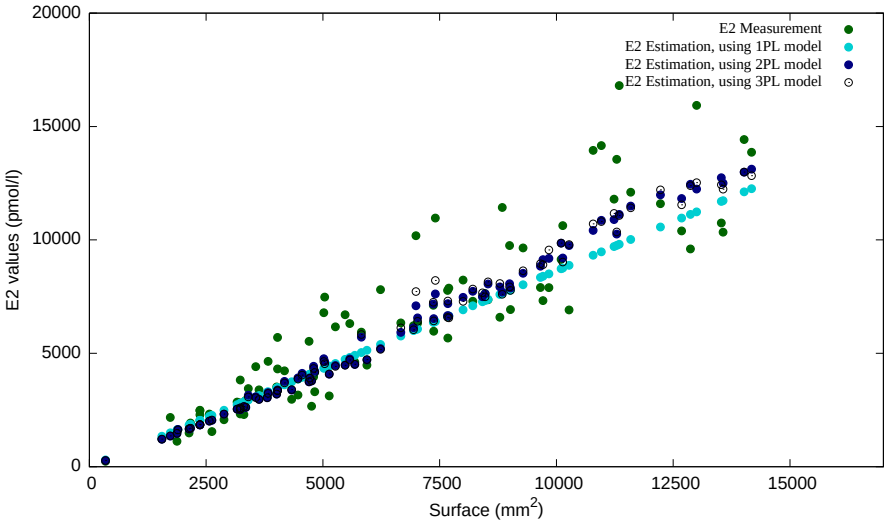

FIGURE B.10: Estimation of E2, for <ER, Other, FSH/LH, Id.1> group, using 1-PL, 2-PL, and 3-PL models.

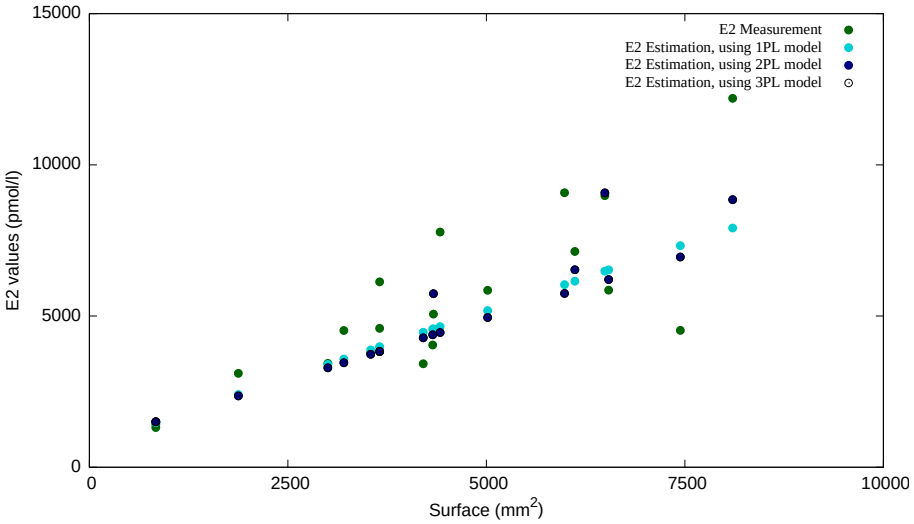

FIGURE B.11: Estimation of E2, for <ER, Other, FSH/LH, Id.2> group, using 1-PL, 2-PL, and 3-PL models.

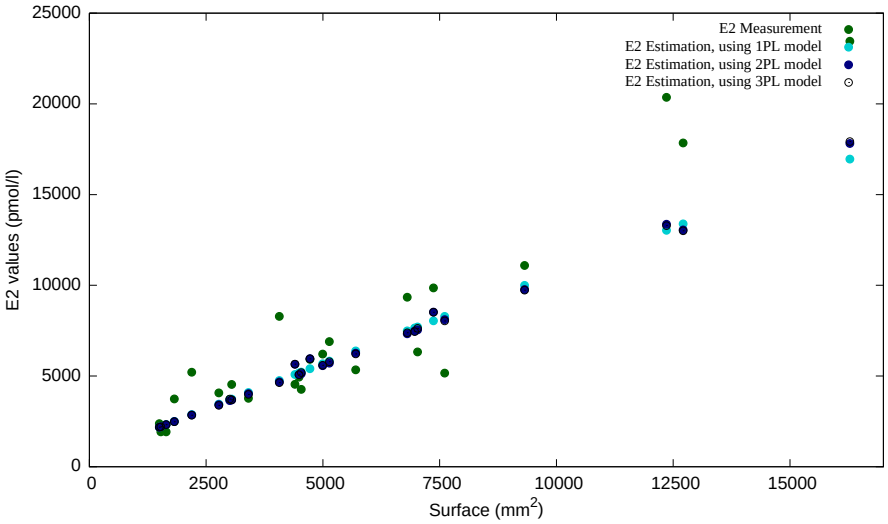

FIGURE B.12: Estimation of E2, for <ER, Other, FSH/LH, Id.3> group, using 1-PL, 2-PL, and 3-PL models.

Appendix C

Graphics from E2 hormone concentration Results Section: Relative % Error obtained by Optimizer build on Relative error in comparison to Relative % Error obtained by Optimizer build on Absolute error.

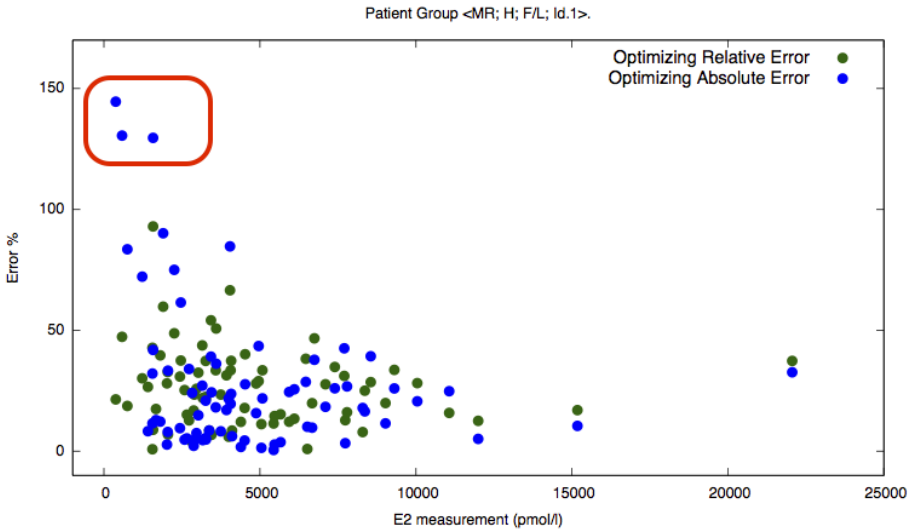

FIGURE C.1: Shows relative percentage error for each measurement of E2 in the <MR, H, F/L, Id.1> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the NPL model.

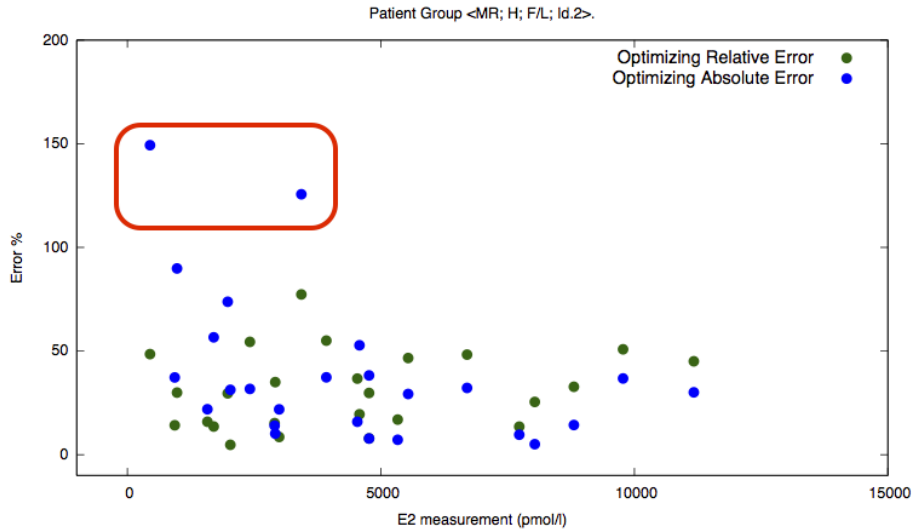

FIGURE C.2: Shows relative percentage error for each measurement of E2 in the <MR, H, F/L, Id.2> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the NPL model.

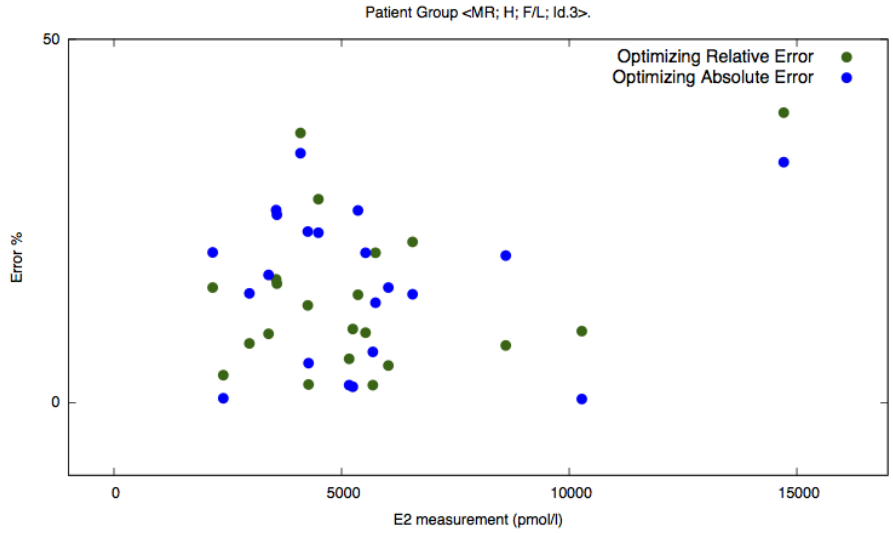

FIGURE C.3: Shows relative percentage error for each measurement of E2 in the <MR, H, F/L, Id.3> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the NPL model.

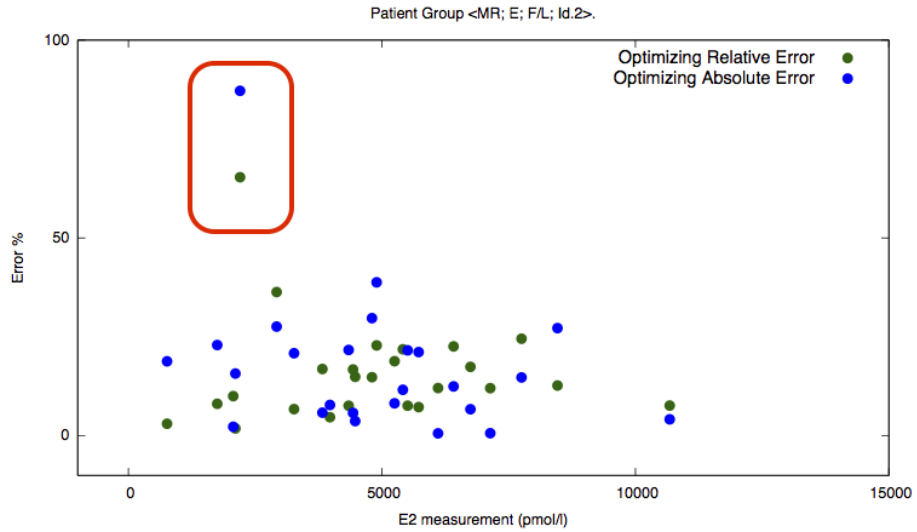

FIGURE C.4: Shows relative percentage error for each measurement of E2 in the <MR, Endo, FSH/LH, Id.2> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the NPL model.

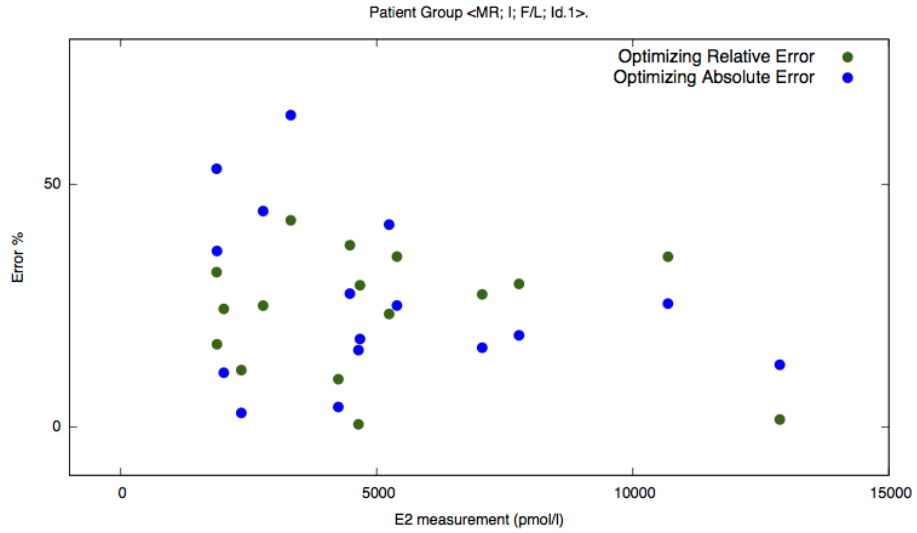

FIGURE C.5: Shows relative percentage error for each measurement of E2 in the <MR, I, F/L, Id.1> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the NPL model.

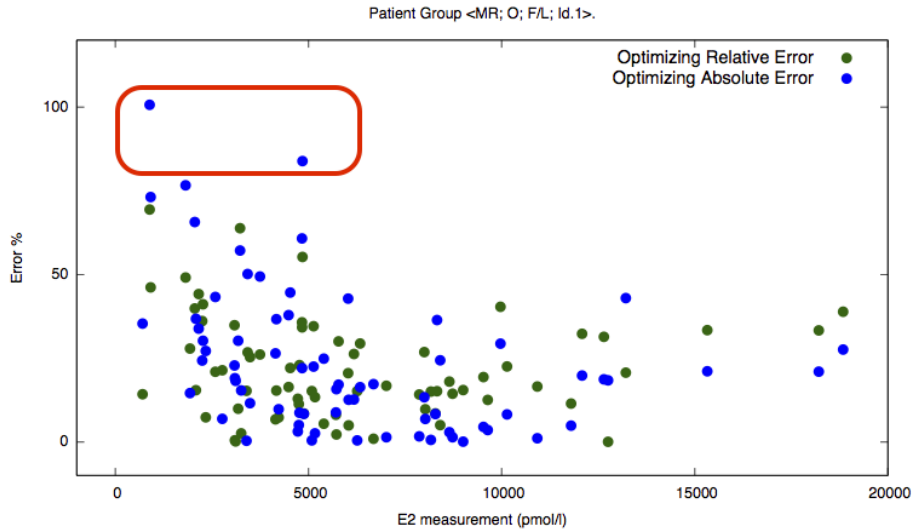

FIGURE C.6: Shows relative percentage error for each measurement of E2 in the <MR, O, F/L, Id.1> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the NPL model.

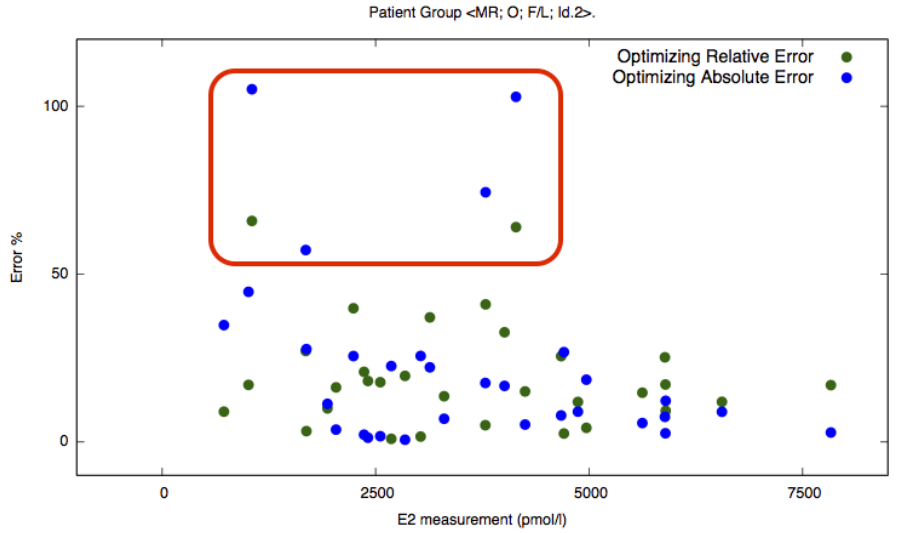

FIGURE C.7: Shows relative percentage error for each measurement of E2 in the <MR, O, F/L, Id.2> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the NPL model.

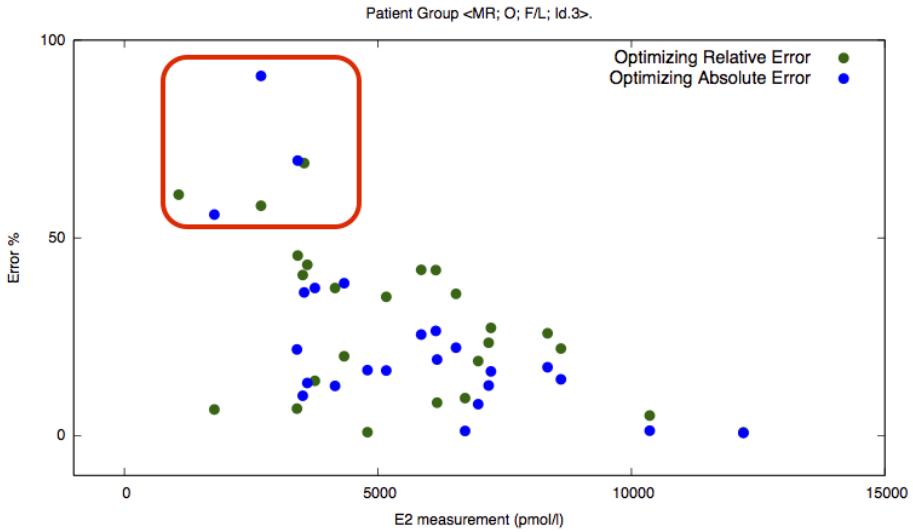

FIGURE C.8: Shows relative percentage error for each measurement of E2 in the <MR, O, F/L, Id.3> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the NPL model.

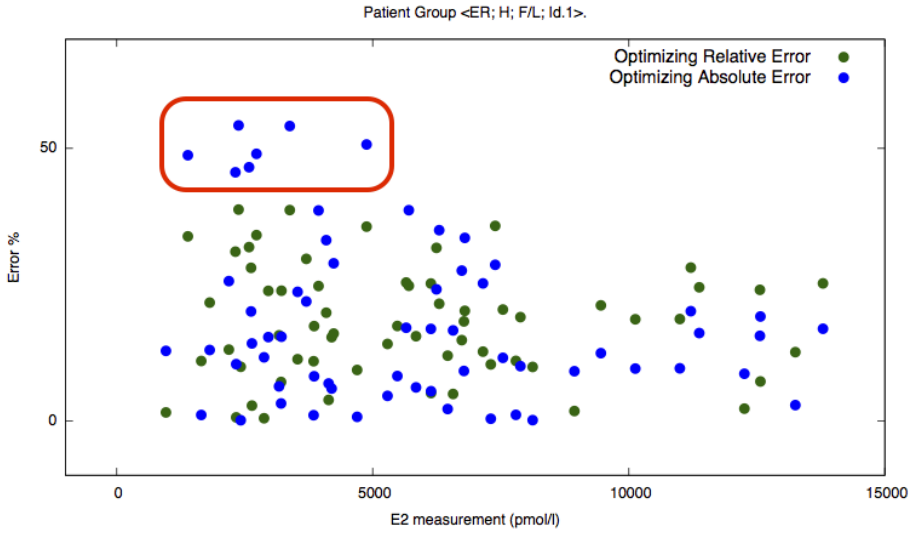

FIGURE C.9: Shows relative percentage error for each measurement of E2 in the <ER, H, F/L, Id1> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the NPL model.

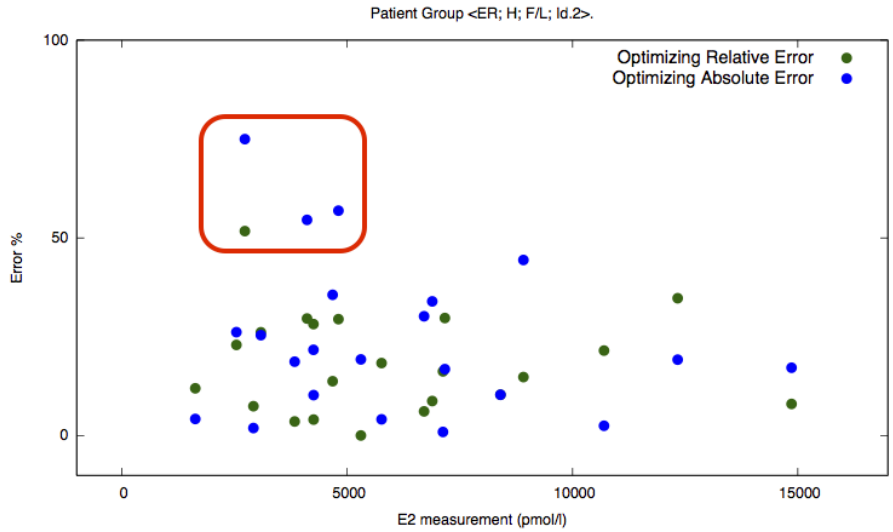

FIGURE C.10: Shows relative percentage error for each measurement of E2 in the <ER, H, F/L, Id2> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the NPL model.

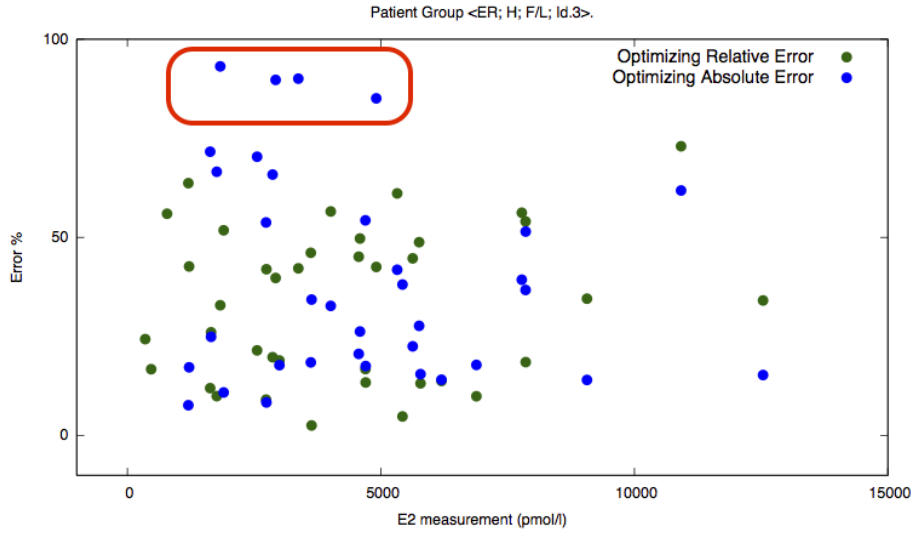

FIGURE C.11: Shows relative percentage error for each measurement of E2 in the <ER, H, F/L, Id3> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the NPL model.

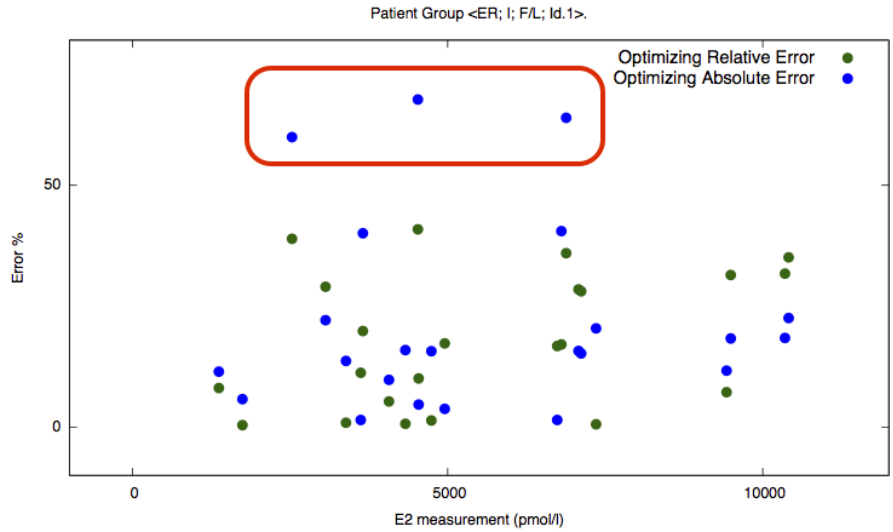

FIGURE C.12: Shows relative percentage error for each measurement of E2 in the <ER, I, F/L, Id1> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the NPL model.

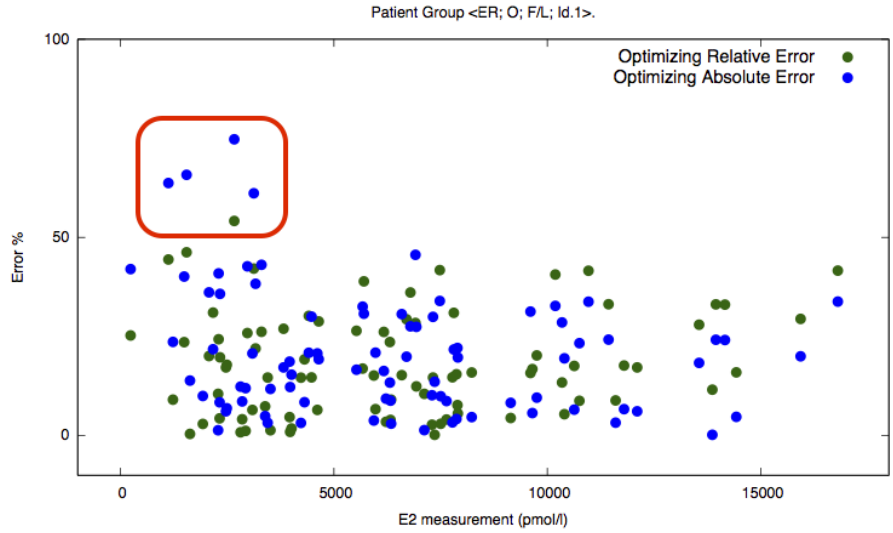

FIGURE C.13: Shows relative percentage error for each measurement of E2 in the <ER, O, F/L, Id1> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the NPL model.

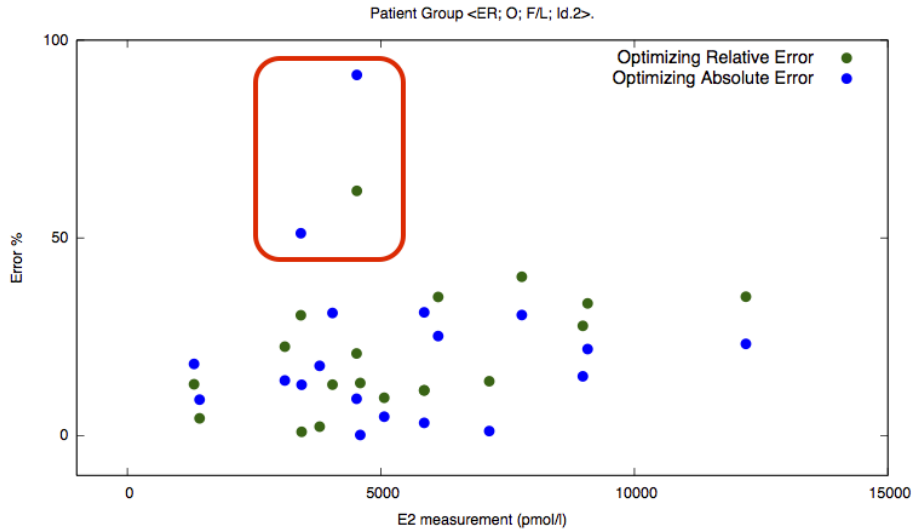

FIGURE C.14: Shows relative percentage error for each measurement of E2 in the <ER, O, F/L, Id2> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the NPL model.

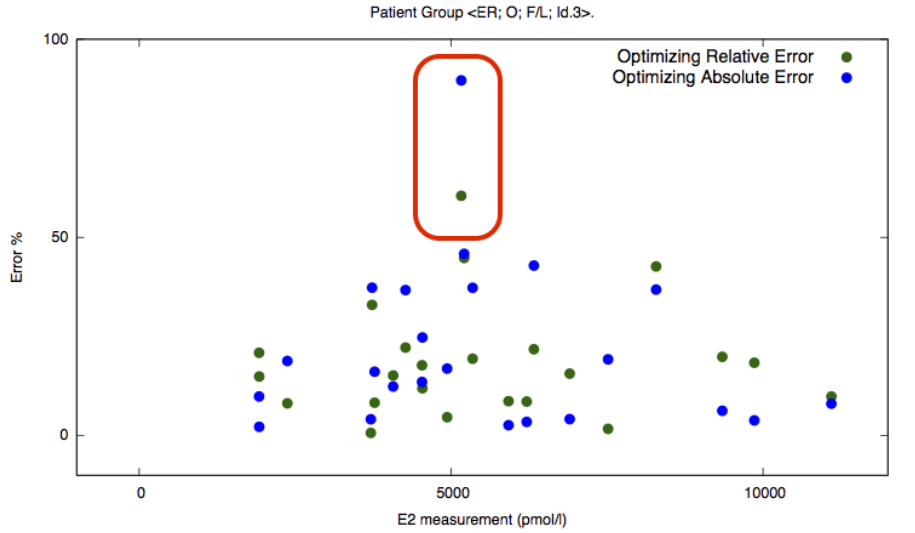

FIGURE C.15: Shows relative percentage error for each measurement of E2 in the <ER, O, F/L, Id3> group, using two optimizers. One optimizing relative error (errors shown by green colour), second optimizing absolute error (errors shown by blue points). Both optimizers are build for the NPL model.

Bibliography

Aboulghar, Mohamed (2003). "Prediction of ovarian hyperstimulation syndrome (OHSS) Estradiol level has an important role in the prediction of OHSS". In: *Human Reproduction* 18.6, pp. 1140–1141.

Al-Hussaini, Tarek K (2012). "OHSS-free IVF practice: Dream or reality". In: *Middle East Fertility Society Journal* 17.1, A1–A3.

Bächler, M et al. (2014). "Species-specific differences in follicular antral sizes result from diffusion-based limitations on the thickness of the granulosa cell layer". In: *Molecular human reproduction* 20.3, pp. 208–221.

Baerwald, Angela R, Gregg P Adams, and Roger A Pierson (2003). "A new model for ovarian follicular development during the human menstrual cycle". In: *Fertility and sterility* 80.1, pp. 116–122.

Clément, Frédérique and Danielle Monniaux (2013). "Multiscale modelling of ovarian follicular selection". In: *Progress in biophysics and molecular biology* 113.3, pp. 398–408.

Clément, Frédérique et al. (2013). "Coupled somatic cell kinetics and germ cell growth: multiscale model-based insight on ovarian follicular development". In: *Multiscale Modeling & Simulation* 11.3, pp. 719–746.

Conover, Cheryl A et al. (2001). "Pregnancy-associated plasma protein-A is the insulin-like growth factor binding protein-4 protease secreted by human ovarian granulosa cells and is a marker of dominant follicle selection and the corpus luteum". In: *Endocrinology* 142.5, pp. 2155–2158.

Daar, Abdallah S and Zara Merali (2002). "Infertility and social suffering: the case of ART in developing countries". In: *Current practices and controversies in assisted reproduction*, pp. 15–21.

D'Angelo, Arianna et al. (2004). "Value of the serum estradiol level for preventing ovarian hyperstimulation syndrome: a retrospective case control study". In: *Fertility and sterility* 81.2, pp. 332–336.

Echenim, Nki et al. (2005). "Multi-scale modeling of the follicle selection process in the ovary". In: *Mathematical biosciences* 198.1, pp. 57–79.

Egli, Marcel, Brigitte Leeners, and Tillmann HC Kruger (2010). "Prolactin secretion patterns: basic mechanisms and clinical implications for reproduction". In: *Reproduction* 140.5, pp. 643–654.

Franco, JG et al. (1993). "Calculation of plasma estradiol levels by analysis of number and size of follicles measured by ultrasound". In: *International Journal of Gynecology & Obstetrics* 41.3, pp. 261–264.

Gerris, Jan and Petra De Sutter (2010). "Self-operated endovaginal telemonitoring (SOET): a step towards more patient-centred ART?" In: *Human reproduction* 25.3, pp. 562–568.

Gnoth, Ch et al. (2003). "Time to pregnancy: results of the German prospective study and impact on the management of infertility". In: *Human Reproduction* 18.9, pp. 1959–1966.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). *The elements of statistical learning 2nd edition*.

Kamel, Remah MA (2013). "Assisted reproductive technology after the birth of louise brown". In: *Gynecology & Obstetrics* 2013.

Karr, Jonathan R et al. (2012). "A whole-cell computational model predicts phenotype from genotype". In: *Cell* 150.2, pp. 389–401.

Kim, In-Cheol and Yong-Gyu Jung (2003). "Using Bayesian networks to analyze medical data". In: *International Workshop on Machine Learning and Data Mining in Pattern Recognition*. Springer, pp. 317–327.

Kwan, Irene et al. (2014). "Monitoring of stimulated cycles in assisted reproduction (IVF and ICSI)". In: *The Cochrane Library*.

Malhotra, Neena (2015). "Endocrine Monitoring of ART Cycles". In: *Principles and Practice of Controlled Ovarian Stimulation in ART*. Springer, pp. 213–221.

Mason, HELEN D et al. (1994). "Estradiol production by granulosa cells of normal and polycystic ovaries: relationship to menstrual cycle history and concentrations of gonadotropins and sex steroids in follicular fluid." In: *The Journal of Clinical Endocrinology & Metabolism* 79.5, pp. 1355–1360.

Mittal, Suneeta et al. (2014). "Serum estradiol as a predictor of success of in vitro fertilization". In: *The Journal of Obstetrics and Gynecology of India* 64.2, pp. 124–129.

Orvieto, Raoul et al. (2007). "The influence of estradiol/follicle and estradiol/oocyte ratios on the outcome of controlled ovarian stimulation for in vitro fertilization". In: *Gynecological endocrinology* 23.2, pp. 72–75.

Orvieto, Raoul et al. (2008). "Controlled ovarian hyperstimulation: are we monitoring the appropriate sex-steroid hormones?" In: *Fertility and sterility* 89.5, pp. 1269–1272.

PAEON (2016a). *Model driven computation of treatments for infertility related endocrinological diseases*. `http://paeon.di.uniroma1.it`. Online; Accessed: 2017-10-20.

— (2016b). *PAEON Project Documentation*. `http://paeon.di.uniroma1.it/docs`. Online; Accessed: 2017-10-20.

Panza, Nicole M, Andrew A Wright, and James F Selgrade (2016). "A delay differential equation model of follicle waves in women". In: *Journal of biological dynamics* 10.1, pp. 200–221.

Papanikolaou, Evangelos G et al. (2006). "Incidence and prediction of ovarian hyperstimulation syndrome in women undergoing gonadotropin-releasing hormone antagonist in vitro fertilization cycles". In: *Fertility and sterility* 85.1, pp. 112–120.

Passmore, Leah et al. (2003). "Assessing decision tree models for clinical in-vitro fertilization data". In: *Dept. of Computer Science and Statistics University of Rhode Island, Technical Report TR03-296*.

Penzias, Alan S et al. (1994). *Ultrasound prediction of follicle volume: is the mean diameter reflective?*

Raine-Fenning, N et al. (2008). "SonoAVC: a novel method of automatic volume calculation". In: *Ultrasound in Obstetrics & Gynecology* 31.6, pp. 691–696.

Röblitz, Susanna et al. (2013). "A mathematical model of the human menstrual cycle for the administration of GnRH analogues". In: *Journal of theoretical biology* 321, pp. 8–27.

Rosendahl, Mikkel et al. (2010). "True ovarian volume is underestimated by two-dimensional transvaginal ultrasound measurement". In: *Fertility and sterility* 93.3, pp. 995–998.

Soboleva, TK et al. (2000). "A model of follicular development and ovulation in sheep and cattle". In: *Animal reproduction science* 58.1, pp. 45–57.

Sohlberg, Björn (1998). "Grey box modelling". In: *Supervision and Control for Industrial Processes*. Springer, pp. 7–43.

Sonaura (2016). *A platform for telemonitoring follicle growth and patient communication*. http://fertihome.com. Online; Accessed: 2016-11-07.

Szmelskyj, Irina and Lianne Aquilina (2014). *Acupuncture for IVF and Assisted Reproduction: An Integrated Approach to Treatment and Management*. Elsevier Health Sciences.

Vandekerckhove, Frank et al. (2014). "Adding serum estradiol measurements to ultrasound monitoring does not change the yield of mature oocytes in IVF/ICSI". In: *Gynecological Endocrinology* 30.9, pp. 649–652.

Var, Turgut et al. (2011). "Relationship between the oestradiol/oocyte ratio and the outcome of assisted reproductive technology cycles with gonadotropin releasing hormone agonist". In: *Gynecological Endocrinology* 27.8, pp. 558–561.

Wang, Xiaobin et al. (2003). "Conception, early pregnancy loss, and time to clinical pregnancy: a population-based prospective study". In: *Fertility and sterility* 79.3, pp. 577–584.