Protocol ITFE-2026-C10

A PHASE II PROSPECTIVE, RANDOMIZED, DOUBLE-BLIND, PLACEBO-CONTROLLED AND MULTI-CENTRE CLINICAL TRIAL TO ASSESS THE SAFETY OF 0.005 % ESTRIOL VAGINAL GEL IN HORMONE RECEPTOR-POSITIVE POSTMENOPAUSAL WOMEN WITH EARLY STAGE BREAST CANCER IN TREATMENT WITH AROMATASE INHIBITOR IN THE ADJUVANT SETTING. "BLISSAFE Study"

STATISTICAL ANALYSIS PLAN

(Version No. 4.0 – 05-Jun-17)

Sponsor: ITF RESEARCH PHARMA S.L.U
Polígono Industrial de Alcobendas
C/San Rafael, 3
28108 Alcobendas (Madrid), Spain
Phone: +34 91 6572323
Fax: +34 91 6572366
Email: italfarmaco@itfsp.com
Protocol-specific Sponsor contact information can be found in the Administrative Binder.

STUDY CHIEF INVESTIGATOR:
Dr. Pedro Sánchez Rovira, Complejo Hospitalario de Jaén
Sponsor Study Code: ITFE-2026-C10
EudraCT Number: 2014-004517-84
Protocol Approved by ITF RESEARCH

Protocol No.: *ITFE-2026-C10 (BLISSAFE) v1.0 26-Nov-2014*

Protocol title: *A phase II prospective, randomized, double-blind, placebo-controlled and multi-centre clinical trial to assess the safety of 0.005% estriol vaginal gel in hormone receptor-positive postmenopausal women with early stage breast cancer in treatment with aromatase inhibitor in the adjuvant setting. "BLISSAFE study"*

Protocol Version: 1.0 26-November-2014

Sponsor: ITF RESEARCH PHARMA S.L.U

This Statistical Analysis Plan was created according the ICH Good Clinical Practice, GEICAM policies and Standard Operating Procedures (SOP); and is consistent with the study protocol.

Version Date: 05-Jun-17

Version No.: 4.0

Previous Version: 3.0

Prepared by Statistician:		
MARIBEL CASAS	Maribel Casas	21-09-2017
Name	Signature	Date
Reviewed by:		
Statistician		
Mª José Escudero		26-09-2017
Name	Signature	Date
Project Manager		
YOLANDA AMIGO		21-09-2017
Name	Signature	Date
Approved by:		
ITF Research Pharma S.L.U. Attorney		
CONCEPCIÓN NIETO		19.9.2017
Name	Signature	Date

Medical Coordinator

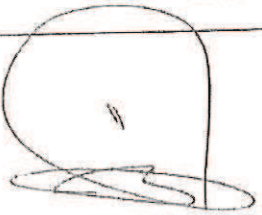
Name	Signature	Date
Pedro Jiménez Rivas		03/08/2017

TABLE OF CONTENTS

ABBREVIATIONS AND DEFINITIONS

AE	Adverse Event
AI	Aromatase Inhibitor
CR	Complete Response
CTCAE	Common Terminology Criteria for Adverse Events
ECOG	Eastern Cooperative Oncology Group
eCRF	Electronic Case Report Form (sometimes referred to as Clinical Report Form). An electronic form for recording study participants' data during a clinical study, as required by the protocol.
ER	Estrogen Receptor
FDA	Food and Drug Administration
FSFI	Female Sexual Function Index
FSH	Follicle Stimulating Hormone
GCP	Good Clinical Practice
GEICAM	Spanish Breast Cancer Research Group
HER2	Human Epidermal Growth Factor Receptor 2
HR	Hormone Receptor
ICD	Informed Consent Document
IDMC	Independent Data Monitoring Committee
IHC	Immunohistochemistry
ITF	ITF Research Pharma S.L.U.
ITT	Intent To Treat
LH	Luteinizing Hormone
MedDRA	Medical Dictionary for Regulatory Activities

MV	Maturation Value
NSAI	Non-Steroidal Aromatase Inhibitor
Patient	A subject with a defined disease
PgR	Progesterone Receptor
QoL	Quality of Life
SAE	Serious Adverse Event
SAP	Statistical Analysis Plan

1. INTRODUCTION

The purpose of this Statistical Analysis Plan (SAP) is to give a detailed description of the statistical analysis to be performed to generate the final study report ITFE-2026-C10 (BLISSAFE) [1]. This includes, on the one hand, a brief summary of the main features of the study, and on the other hand the objective of SAP corresponding to the planning of the statistical analysis of the study.

This Statistical Analysis Plan was created according the ICH Good Clinical Practice [2] [3], GEICAM policies and Standard Operating Procedures (SOP); and is consistent with the study protocol [1].

2. STUDY OBJECTIVES

2.1 Main Objective

The primary objective is to evaluate the levels of FSH after treatment with 0.005% estriol vaginal gel in hormone receptor-positive postmenopausal women with early stage breast cancer in treatment with NSAIs in the adjuvant setting and symptoms of vaginal atrophy.

2.2 Secondary Objective

- To evaluate the levels of estriol, estradiol, estrone, FSH and LH after treatment with 0.005% estriol vaginal gel in hormone receptor-positive postmenopausal women with early stage breast cancer in treatment with NSAIs in the adjuvant setting and symptoms of vaginal atrophy.
- To assess the safety and tolerability of 0.005% estriol vaginal gel in hormone receptor-positive postmenopausal women with early stage breast cancer in treatment with NSAIs in the adjuvant setting and symptoms of vaginal atrophy.
- To assess the efficacy of 0.005% estriol vaginal gel in the treatment of symptoms and signs of vaginal atrophy in hormone receptor-positive postmenopausal women with early stage breast cancer in treatment with NSAIs and symptoms of vaginal atrophy.

- To measure the impact of treatment with 0.005% estriol vaginal gel in sexual function of hormone receptor-positive postmenopausal women with early stage breast cancer in treatment with NSAIs and symptoms of vaginal atrophy.

3. STUDY DESIGN

This is a phase II, prospective, randomized, double-blind, placebo-controlled, international (Spain and Sweden) and multicentre study.

60 patients will be randomized to receive 0.005% estriol vaginal gel (arm A) and placebo moisturizing gel (arm B) in a 4:1 proportion, so 48 patients will enter in arm A and 12 in arm B.

Arm A (experimental arm): Investigational drug product: 0.005% estriol vaginal gel (Blissel®)

Route:	Vaginal. Administration by an applicator inserted deep inside the vagina
Dose:	1 g of gel, containing 50 μg of estriol
Dosage schedule:	Weeks 1-3: single daily application Weeks 4-12: twice weekly administration.

Women will be instructed to administer the gel at bedtime.

Arm B (control arm): placebo moisturizing gel (vehicle of Blissel®) administered in the same way.

Patients will have hormone determinations (estriol, estradiol, estrone, FSH and LH) analyzed in a central laboratory at baseline and at weeks 1, 3, and 8 from the beginning of the therapy and at week 12 (last day of study therapy). An additional FSH and LH determination will be obtained during the screening period in order to assess the intra-individual variation of these hormones in baseline conditions. Patients will also undergo gynecological examination (vaginal smear and pH), evaluation of vaginal symptoms and signs, and will answer a female sexual function index questionnaire at baseline, at week 3 (from the beginning of therapy) and week 12 (last day of study therapy). Adverse events will be followed along the study and specifically collected at weeks 1, 3 and 8 from the beginning of therapy, at week 12 (last day of study therapy) and at safety visit (30 ±5 days after last day of study therapy).

There will be an initial safety phase of the study (only two sites of Spain will participate), in which 10 patients will be included (also randomized 4:1) to obtain some preliminary safety data and FSH

values prior to the recruitment of the rest of the group under study. These women will receive the study treatment over 3 weeks (8 will receive 0.005% estriol vaginal gel and 2 placebo moisturizing gel). FSH levels will be analyzed at screening, baseline and after 3 weeks and reviewed (together with safety data) by an independent data monitoring committee (IDMC) which will advise the chief investigator about a go /no-go decision with the other 60 patients to complete the calculated sample size. The study will continue if no significant changes in FSH levels from postmenopausal to premenopausal values are demonstrated.

3.1 Sample Size

This is an exploratory study and there are scarce data to allow a precise sample size calculation. The most relevant data come from Pfeiler [4] who reported that the levels of FSH in patients treated with aromatase inhibitors fell from a mean value of 75.7 to 66.0 mIU/ml after two weeks of daily vaginal treatment with 0,5 g of estriol. Although no figures are reported of the variability of those measurements, an estimation of the standard deviation can be made applying the standard rule of [sd = range/4]. The reported range of FSH values in that report is 45.6 – 134.6, thus a reasonable approximation to the sd is 89/4 = 22.3. Under these assumptions, it can be calculated that a sample size of 44 would provide 80% power to detect with an alpha=0.05 a decrease of FSH levels from 75.7 to 66.0 mIU/ml assuming a standard deviation of 22.3.

The hormone levels in women treated with AI who are not receiving estrogen therapy are likely to suffer physiological temporal oscillations. These variations could make difficult the interpretation of any eventual modification of hormone levels observed in the estriol treated women. In the absence of data of naturally occurring variations in hormone levels of women treated with AI that allow us to interpret the results of hormone determinations under study, it is considered appropriate to include a placebo group to provide reference data. In this case, the placebo is a vaginal moisturizing gel.

It is estimated that a group of 11 women will allow the assessment of physiological fluctuations in hormone levels. Additionally, this sample size for the placebo group will provide 80% power and alpha=0.05 to differentiate between active and placebo (ratio 4:1) on the change in the vaginal maturation value, which is a secondary variable of the study that provides objective information about the effectiveness of treatment on vaginal atrophy (calculation based on the parameter data from the pivotal Blissel study and considering only the subgroup of women who had severe vaginal dryness: change MV active group 24.9 vs 0.6 points in placebo group, SD = 24.2).

The differentiation between active and placebo in the subjective parameters of efficacy (secondary) would be possible at the expense of a sample size noticeably superior. Since the evaluation of effectiveness is not primary but secondary in this study, it is not justified in this study to increase the sample for this reason.

In accordance with the above, the study would include 55 women. 44 will receive vaginal estriol gel and 11 with the placebo moisturizing gel). Taking into account 10% of dropouts, the sample to be considered is of 60 women (48 vaginal estriol gel and 12 placebo moisturizing gel).

10 additional women will be recruited in a first safety phase of the study (8 with vaginal estriol gel and 2 placebo moisturizing gel).

3.2 Randomization

This is an unbalanced study but with an active: placebo ratio of 4:1, which will support an adequate patient randomization (blocks of 5).

This is a double blind study, the study drug, 0.005% Estriol vaginal gel, and its vehicle in gel (placebo vaginal gel) will be of identical appearance and will have the same texture in order to maintain the double blind.

The medication will be randomized at ITF Research Pharma according to a randomization list performed at ITF Research Pharma. The medication will be sent in blocks of 5 to each site, and the first patient included in a specific site, will receive the lowest randomization code indicated in the medication received. Codes will be assigned consecutively from the lowest number to the highest. ITF provides us and also each site, with opaque and sealed code break envelopes, identified with the randomization code, and containing inside a card with the treatment assigned to that patient. At the end of the study, and after the clinical Data Base has been closed, ITF Research Pharma will send us the complete randomization list.

For more information about the unblinding process please see the protocol (part 5.8).

4. STUDY POPULATION

Patients with hormone receptor positive and any HER2 status early breast cancer, who are on treatment with NSAI and have vaginal atrophy.

4.1 Inclusion Criteria

Patients are eligible to be included in the study only if they **meet all** of the following criteria:

1. Written informed consent prior to beginning specific protocol procedures.
2. Patients must have histological confirmation of breast adenocarcinoma with stage I-IIIA, documented at a local pathology department.
3. The breast tumors must be estrogen-receptor positive and/or progesterone receptor positive (≥1% of stained tumor cells by IHC as determined by the local laboratory) with any HER2 status.
4. Postmenopausal status defined as: 12 months of spontaneous amenorrhea or 6 months of spontaneous amenorrhea with serum FSH levels > 40 mIU/ml or 6 weeks postsurgical bilateral oophorectomy with or without hysterectomy.
5. Patient must be receiving the non-steroidal aromatase inhibitors anastrozole or letrozole as breast cancer treatment in the adjuvant setting for a minimum of 6 months.
6. Women suffering from moderate to severe vaginal dryness according to the FDA guidelines for drug development in postmenopausal women (Center for Drug Evaluation and Research, CDER Jan 2003). A moderate symptom will be considered if the symptom is present, bothersome and annoying, and a severe symptom will be considered if the symptom is present, bothersome and annoying, and interferes with the normal patient activity.
7. Eastern Cooperative Oncology Group Performance Status (ECOG PS) of 0 or 1.
8. Adequate bone marrow as defined by the following laboratory values:
    a. Absolute Neutrophil Count (ANC) ≥ 1.5 x 10^9/L.
    b. Platelets (plt) ≥ 100 x 10^9/L.
    c. Hemoglobin (Hgb) ≥ 10 g/dl.
9. Patient has adequate organ function as defined by the following laboratory values:
    d. Serum creatinine ≤ 1.5 x ULN.
    e. Bilirubin ≤ 1.5 × ULN.

f. Alkaline phosphatase ≤ 2 × ULN.

g. AST and ALT ≤ 2 × ULN.

10. Willingness and ability to comply with scheduled visits, treatment plan, laboratory tests and other study procedures.

4.2 Exclusion Criteria

Patients will be excluded from the study if they **meet any** of the following criteria:

1. Stage IIIB-IV breast cancer or bilateral breast cancer.
2. Treatment with any other current anti-tumoral therapy (chemotherapy, anti-Her2…etc) besides the NSAI. Pamidronate or Alendronate are permitted.
3. Prior history of other malignancy within 5 years of study entry, aside from non-melanoma skin cancer or carcinoma-in-situ of the uterine cervix adequately treated.
4. Postmenopausal uterine bleeding. Vaginal bleeding of unknown etiology.
5. Patients with endometrial thickness equal to or greater than 4 mm measured by transvaginal ultrasound.
6. Patients who have received any type of vulvovaginal treatment in the 15 days prior to the start of the study.
7. Use of any hormone, natural (phytoestrogens) or herbal products for the treatment of menopausal symptoms within the last 3 months.
8. Current or previous history of thromboembolic disease or coagulopathies.
9. Severe cardiovascular or respiratory diseases in the previous 6 months.
10. Renal Impairment.
11. Hepatitis B and/or hepatitis C carriers (unless with normal hepatic function).
12. Known human immunodeficiency virus infection.
13. Known hypersensitivity to NSAI.

14. Other severe acute or chronic medical or psychiatric condition, or laboratory abnormality that would impart, in the judgment of the investigator, excess risk associated with study participation or study drug administration, or which, in the judgment of the investigator, would make the patient inappropriate for entry into this study.
15. Previous investigational treatment for any condition or participation in any clinical trial within 4 weeks of inclusion date.

4.3 Discontinuations

4.3.1 Discontinuation of Study Medication

The criteria for enrollment must be followed explicitly. If a patient who does not meet enrollment criteria is inadvertently enrolled, that patient should be discontinued from the study medication, an exception may be granted if the patient, in the opinion of the investigator, is having benefit from the study medication. In these rare cases, the investigator must obtain documented approval from the sponsor to allow the patient to continue to receive the study drug.

Patients can be discontinued from the study therapy in the following circumstances:

- Patient's own request.
- Unacceptable toxicity.
- Tumor recurrence.
- Any clinical adverse event (AE), laboratory abnormality or inter-current illness which, in the opinion of the investigator, indicates that continued participation in the study is not in the best interest of the patient.
- Termination of the study by the sponsor.
- Physician's decision, including need of other anti-cancer therapy, not specified in the protocol.
- If the patient is non-compliant with study procedures.
- Loss of ability to freely provide consent through imprisonment or involuntarily incarceration for treatment of either a psychiatric or physical (e.g. infectious disease) illness.

- All patients will be discontinued from the treatment phase and they will perform the safety visit in case of a delay of more than 1 week during the daily administration phase or more than 2 weeks during the twice weekly administration phase, or permanent discontinuation of the study drug unless there is an obvious clinical benefit per the investigator's medical judgment and after discussion with the sponsor.

If possible, and after the permanent discontinuation of treatment, the patients will be assessed using the procedure normally planned for the last dosing day with the study treatment.

All permanent treatment discontinuation should be recorded by the Investigator in the eCRF when considered as confirmed.

4.4 ITT Population

The intent to treat (ITT) population will include all patients who are randomized, with study treatment (Estriol gel vaginal 0.005% or placebo) assignment designated according to initial randomization.

The ITT population will be the primary population for evaluating patient characteristics, hormone levels, efficacy and the FSFI questionnaire.

4.5 Per protocol population

Per-protocol population is a subset of the ITT population that received at least one dose of study drug and completed the study without protocol violations according to Protocol Deviation Manual.

4.6 Safety population

Safety population will include all patients randomized in the study who received at least one dose of treatment, and they will be analyzed according to the actual treatment received. This population is for the safety analysis.

5. ENDPOINTS AND STUDY VARIABLES

5.1 Primary Endpoint

Variation in serum levels of FSH from baseline to 12 weeks of treatment.

5.2 Secondary Endpoints

- Variation in serum levels of FSH at different time points compared to baseline (weeks 1, 3 and 8).
- Variation in serum levels of LH and plasma levels of estriol, estradiol and estrone, at different time points compared to baseline (weeks 1, 3, 8 and 12).
- AEs according to the Medical Dictionary for Regulatory Activities (MedDRA).
- Changes in vaginal dryness and other symptoms and signs of vaginal atrophy; changes in vaginal maturation value and changes in vaginal pH at week 3 and week 12 vs baseline.
- Changes in sexual function measured by the Female Sexual Function Index (FSFI [5]) scale at week 3 and week 12 vs baseline.

6. DATA SCREENING AND ACCEPTANCE

6.1 Missing data

The frequency of missing data will be examined and reported for each variable in the analysis. Missing data for FSH, LH, Estriol, Estradiol and Estrone values will be imputed with the last observation carried forward (LOCF) method. Analyses without data imputation will be performed to compare the results. For adverse events and rest of study variables, we will not perform data imputation for missing data.

6.1.1 Missing date

All non-relevant missing dates that cannot be determined by query should be estimated as follow:

If the day of the month is missing for any date used in a calculation, the 15th of the month will be used to replace the missing date unless the calculation results in a negative time duration (e.g., date of onset cannot be prior to day one date). In this case, the date resulting in 1 day

duration will be used. If the day of the month and the month are missing for any date used in a calculation, the date will be considered missing.

There will be performed a list with all the date replaced and it will be reviewed by another statistician different from the study statistician.

6.2 Statistical software

The statistical analysis will be developed in SAS Enterprise Guide v5.1.

6.3 Database lock

-The data base lock for the interim analysis is planned after the first 10 patients have been included and the results of the FSH levels at week 3 will be available and the patient have completed the safety visit 30(+/-5) days after the last study drug administration and all the queries for those 10 patients have been closed. The randomization codes will be unblinded and made available for the project statistician to perform the data analysis, when the study for those 10 patients has been finalized, the database has been verified and the protocol violations have been determined. The data base must be opened again only if there is a mistake affecting a principal variable than can affect the integrity of the efficacy or safety results.

-The database lock for the final analysis will occur approximately 6 months from the inclusion of the last patient, if the study meets the plan set. For the database lock all the patients must have been performed the safety visit and all the queries must be closed. The randomization codes will be unblinded and made available for the project statistician to perform the data analysis, when the study has been finalized, the database has been verified and the protocol violations have been determined. The data base must be opened again only if there is a mistake affecting a principal variable than can affect the integrity of the efficacy or safety results.

7. INTERIM ANALYSIS

7.1 Purpose of interim analysis

There will be an interim analysis after the first 10 included patients have completed 3 weeks of therapy. These patients will be evaluated after 3 weeks according to the study scheme.

Only safety data and the analytical values of FSH of these 10 first patients will be reviewed. It is expected that no significant changes from the postmenopausal to the premenopausal range occurs in any patient.

According to the reference range provided by the central laboratory for the FSH determinations (Laboratorios Echevarne), it is shown in the table below, that 95% of postmenopausal women have values from 21.7 to 153. A value of 22 mIU/ml will be set as long as baseline value will be greater than this value. In any case, all patients will be evaluated by the IDMC.

		FSH, mIU/ml	
Group	n	Median	Central 95%
Men	135	3,8	0,7 – 11,1
Women			
Postmenopausal	76	90,5	21,7 – 153
Postmenopausal (ERT)	16	27	9,7 – 111
Oral Contraceptives	12	1,7	ND – 4,9
*Preliminary		ND: not detectable	

The 10 patients included in the interim analysis will not be included in the final analysis, as they will not complete 3 months of therapy. For that reason, it will not be necessary to perform a correction of the final alpha.

7.2 IDMC (Interim Data Monitoring Committee)

The study will use an IDMC. The IDMC membership and governance is outlined in a separate charter.

The IDMC will be responsible to review the FSH and safety data from the first 10 patients included in the study and to advise the sponsor and chief investigator about a go /no-go decision with the other 60 patients to complete the calculated sample size.

7.2.1 IDMC Recommendation

After review of interim analysis by the IDMC, it has been requested to record the time of the last administration of Blissel prior to any hormonal extraction to explain possible drastic changes that could occur in the hormonal values and if these changes are due to variation between screening and basal values or to Blissel administration

.

8. CRITERIA FOR END OF STUDY

This study will be considered complete following the data cut-off date and datalock for the final analysis. The data cut-off date for the final analysis will occur after all enrolled patients have completed the safety visit 30 (+/-5) days after the last study drug administration.

9. STATISTICAL METHODS AND ANALYSIS

9.1 Statistical Methods

Descriptive Analysis

For categorical variables it will be calculated frequencies and percentages. For continuous variables it will be calculated standard descriptive statistics, such as total number of observations, number of available data, mean, standard deviation, minimum, percentil 25, median, percentil 75 and maximum.

Chi-square χ^2 Test

The first type of chi-square test is *the goodness of fit test*. This is a test which makes a statement or claim concerning the nature of the distribution for the whole population. The data in the sample is examined in order to see whether this distribution is consistent with the hypothesized distribution of the population or not. One way in which the chi-square goodness of fit test can be used is to examine how closely a sample matches a population. The chi-square goodness of fit test can be used to provide a test for the representativeness of a sample.

Suppose that a variable has a frequency distribution with *k* categories into which the data has been grouped. The frequencies of occurrence of the variable, for each category, are called the observed values. The manner in which the chi-square goodness of fit test works is to determine how many cases there would be in each category if the sample data were distributed exactly according to the claim. These are termed the expected number of cases for each category. The total of the expected number of cases is always made equal to the total of the observed number of cases. The null hypothesis is that the observed number of cases in each category is exactly equal to the expected number of cases in each category. The alternative hypothesis is that the observed and expected number of cases differ sufficiently to reject the null hypothesis.

Let O_i is the observed number of cases in category *i* and E_i is the expected number of cases in each category, for each of the *k* categories *i = 1, 2, 3…., k*, into which the data has been grouped. The hypotheses are

$$H_0 : O_i = E_i \qquad \forall\, i$$
$$H_1 : O_i \neq E_i \quad for\ some\ i$$

and the test statistic is

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

where the summation proceeds across all *k* categories, and it has *k-1* degrees of freedom.

The second type of chi-square test is the chi-square *test for independence of two variables*. The chi-square test of independence allows researchers to determine whether variables are independent of each other or whether there is a pattern of dependence between them. If there is dependence, the researcher can claim that the two variables have a statistical relationship with each other.

The only limitation on the use of this test is that the sample sizes must be sufficiently large to ensure that the expected number of cases in each category is five or more, if not, the Fisher exact test must be used. This rule can be modified somewhat, but as with all approximations, larger sample sizes are preferable to smaller sample sizes. There are no other limitations on the use of the test, and the chi-square statistic can be used to test any contingency or cross classification table for independence of the two variables.

The chi-square test for independence is conducted by assuming that there is no relationship between the two variables being examined. The alternative hypothesis is that there is some relationship between the variables.

$$\mathrm{H}_0 : \mathrm{X\ and\ Y\ are\ independent}$$
$$\mathrm{H}_1 : \mathrm{X\ and\ Y\ \ are\ dependent}$$

Using that we can estimate the expected frequencies in each cell as $E_{ij} = np_ip_j$ where *n* is the total count, p_i is the proportion in row *i* and p_j is the proportion in row *j*. Then the chi-quared statistic is, for an *r* ×*c* table:

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2_{(r-1)(c-1)}$$

The observed numbers of cases, O_{ij}, are the numbers of cases in each cell of the cross classification table, representing the numbers of respondents for each combination of the variables. The expected numbers of cases E_i for each of the cell can be obtained from the multiplication rule of probability for independent events.

Fisher Exact Test

When the assumptions for the Chi-squared test are not met, i.e. one of the expected values in a 2x2 table is less than 5, and especially when it is less than 1, then Fisher's Exact test must be applied. The null hypothesis for the test is that there is no association between the rows and columns of the 2x2 table, such that the probability of a subject being in a particular row is not influenced by being in a particular column. If the columns represent the study group and the rows the outcome, then the null hypothesis could be interpreted as the probability of having a particular outcome not being influenced by the study group, and the test evaluate whether the two study groups differ in the proportions with each outcome. An important assumption for Fisher's Exact test, is that the binary data are independent. If the proportions are correlated then more advanced techniques should be applied.

The test is based upon calculating directly the probability of obtaining the results that we have shown (or results more extreme) if the null hypothesis is actually true, using all possible 2×2 tables that could have been observed, for the same row and column totals as the observed data. These row and column totals are also known as marginal totals. What we are trying to establish is how extreme our particular table (combination of cell frequencies) is in relation to all the possible ones that could have occurred given the marginal totals.

For tables *r*x*k*, with *r* or/and *k* major than 2, it could be used the extension of Freeman-Halton to the Fisher Exact Test. Other alternative is to obtain the p-value using Monte Carlo simulation [6].

Mann-Whitney-Wilcoxon Test

Mann Whitney U o Wilcoxon rank-sum test

The Mann-Whitney U test is a nonparametric method for comparing two independent samples. This test is also called the Wilcoxon rank-sum test. The Mann-Whitney U test assumes only that we have independent random samples from the two groups. It does not assume anything about a normal distribution.

The null hypothesis for the Mann-Whitney U test is that the population distribution of the response variable—whatever that distribution might be— is the same for both groups. The alternative hypothesis is that the response variable tends to be larger for one group than for the other group.

To calculate the test statistic for the Mann-Whitney U test, we start by summing the ranks in each group. Define

R_1 = sum of ranks for Group 1; R_2 = sum of ranks for Group 2:

Then just pick one group or the other—it really doesn't matter which. Let's say we pick Group 1. Then the Mann-Whitney U test statistic[1] is

$$U = R_1 - \frac{n_1(n_1 + 1)}{2}$$

where n_1 is the number of subjects in Group 1.

If the null hypothesis is true, then every possible arrangement of the ranks among the two groups is equally likely. We can use this fact to calculate the probability of getting various values of the test statistic U under the null hypothesis, which in turn lets us calculate the p-value of the test.

If the p-value is less than or equal to α, then we reject the null hypothesis and conclude that one group tends to have larger response variable values than the other (one-sided alternative) or merely that there is a difference one way or the other (two-sided alternative).

If the p-value is greater than α, then we fail to reject the null hypothesis, which means we think the null hypothesis is reasonable, which means we think it's reasonable that there's no difference between the response variable values for the two groups.

[1] There are a lot of other ways to calculate U, all of which give the same answer.

Wilcoxon signed-ranks test

Wilcoxon signed-ranks test is a non-parametric test for correlated samples. The rationale of the test is similar to that of the Mann-Whitney test for independent samples. This test assumes that it can rank differences between paired observations. The data for the test will consist of a number of pairs of scores, each derived from a single subject, or from a pair of matched subjects.

The Wilcoxon signed ranks test requires that the differences are approximately symmetric and that the data are measured on an ordinal, interval, or ratio scale. When the assumptions for the Wilcoxon signed ranks test are met but the assumptions of the t test are violated, the Wilcoxon signed ranks test is usually more powerful in detecting a difference between the two populations.
The Wilcoxon signed ranks test uses the test statistic W computing as follow

1. For each item in a sample of n items, compute a difference score, D_i between the two paired values.
2. Neglect the + and - signs and list the set of n absolute differences, $|\mathrm{D_i}|$
3. Omit any absolute difference score of zero from further analysis, thereby yielding a set of n' nonzero absolute difference scores, where n'< n. After you remove values with absolute difference scores of zero, n' becomes the actual sample size.
4. Assign ranks $\mathrm{R_i}$ from 1 to n' to each of the $|\mathrm{D_i}|$ such that the smallest absolute difference score gets rank 1 and the largest gets rank n'. If two or more$|\mathrm{D_i}|$ are equal, assign each of them the mean of the ranks they would have been assigned individually if ties had not occurred in the data.
5. Reassign the symbol + or - to each of the n' ranks,$\mathrm{R_i}$, depending on whether $\mathrm{D_i}$ was originally positive or negative.
6. Compute the Wilcoxon test statistic, W, as the sum of the positive ranks

$$\mathrm{W} = \sum_{\mathrm{i}=1}^{\mathrm{n}'} \mathrm{R_i^{(+)}}$$

The Wilcoxon signed rank test rejects the hypothesis that there are no systematic differences within pairs when the rank sum W+ is far from its mean.

ANCOVA

The analysis of covariance (ANCOVA) is typically used to adjust or control for differences between the groups based on another, typically interval level, a variable called the covariate. The ANCOVA is an extension of ANOVA that typically provides a way of statistically controlling the effects of continuous or scale variables that you are concerned about, but that are not the focal point or independent variable(s) in the study. The ANCOVA F test evaluates whether the population means on the dependent variable, adjusted for differences on the covariate, differ across levels of a factor. If a factor has more than two levels and the F is significant, follow-up tests should be conducted to determine where there are differences on the adjusted means between groups.

One-way analysis of covariance model

$$y_{ij} = \mu + \tau_i + \beta\,(x_{ij} - \bar{x}..) + \varepsilon_{ij} \quad i = 1,\ldots,k\ ;\ j = 1,\ldots, n_i$$

μ grand mean

τ_i effect of group i.

β lineal regression coefficient.

x_{ij} covariate value corresponding to observation y_{ij}.

$\bar{x}..$ mean of covariate.

ε_{ij} random error.

n_i: number of subjects in treatment i.

The null hypothesis of ANCOVA is no difference among the adjusted population means.

$$\begin{matrix} H_0 : \tau_i = \tau_k \\ H_a : \tau_i \neq \tau_k \end{matrix} \quad for\ some\ i,k$$

But it is necessary to test the following hypothesis previously.

$$\begin{cases} H_0 : \beta = 0 \\ H_a : \beta \neq 0 \end{cases}$$

If the covariate has not predictive power for dependent variable (null hypothesis) we will use ANOVA instead of ANCOVA because it would be exactly the same statistical test.

ASSUMPTIONS FOR ANCOVA

There are two major assumptions that underlie the use of ANCOVA; both concern the nature of the relationship between the dependent variable and the covariate.

The first is that the relationship is linear. If the relationship is nonlinear, the adjustments made in the ANCOVA will be biased; the magnitude of this bias depends on the degree of departure from linearity, especially when there are substantial differences between the groups on the covariate. Thus it is important for the researcher, in preliminary analyses, to investigate the nature of the relationship between the dependent variable and the covariate (by looking at a scatter plot of the data points), in addition to conducting an ANOVA on the covariate.

The second assumption is about the regression lines within each of the groups. We assume that the relationship must be linear. Additionally, however, the regression lines for these individual groups are assumed to be parallel; in other words, they have the same slope. This assumption is often called homogeneity of regression slopes or parallelism and is necessary in order to use the pooled within-groups regression coefficient for adjusting the sample means; this is one of the most important assumptions for the ANCOVA. Failure to meet this assumption implies that there is an interaction between the covariate and the treatment. This assumption can be checked with an F test on the interaction of the independent variable(s) with the covariate(s). If the F test is significant (i.e., significant interaction) then this assumption has been violated and the covariate should not be used as it is. A possible solution is converting the continuous scale of the covariate to a categorical (discrete) variable and making it a subsequent independent variable, and then use a factorial ANOVA to analyze the data.

The assumptions underlying the ANCOVA had a slight modification from those for the ANOVA, however, conceptually, they are the same.

Assumption 1: The cases represent a random sample from the population, and the scores on the dependent variable are independent of each other, known as the *assumption of independence*. The test will yield inaccurate results if the independence assumption is violated. This is a design issue that should be addressed prior to data collection. Using random sampling is the best way of ensuring that the observations are independent; however, this is not always possible. The most important thing to avoid is having known relationships among participants in the study.

Assumption 2: The dependent variable is normally distributed in the population for any specific value of the covariate and for any one level of a factor (independent variable), known as the assumption of normality. This assumption describes multiple conditional distributions of the dependent variable, one for every combination of values of the covariate and levels of the factor, and requires them all to be normally distributed. To the extent that population distributions are not normal and sample sizes are small, p values may be invalid. In addition, the power of ANCOVA tests may be reduced considerably if the population distributions are non-normal and, more specifically, thick-tailed or heavily skewed. The *assumption of normality* can be checked with skewness values (e.g., within +3.29 standard deviations).

Assumption 3: The variances of the dependent variable for the conditional distributions are equal, known as the *assumption of homogeneity of variance*. To the extent that this assumption is violated and the group sample sizes differ, the validity of the results of the one-way ANCOVA should be questioned. Even with equal sample sizes, the results of the standard post hoc tests should be mistrusted if the population variances differ. The assumption of homogeneity of variance can be checked with the Levene's F test.

If the assumptions are not met then it will be applied nonparametric ANCOVA [7]. For example, Akritas et al [8] proposed a completely nonparametric model that avoids restrictive modeling assumptions.

9.2 Statistical Analysis

Standard descriptive statistics, such as the mean, median, range and proportion, will be used to summarize the patient sample and to estimate parameters of interest. Ninety-five percent confidence intervals will be provided for estimates of interest wherever possible.

The variation of the levels of FSH, estriol, estradiol, estrone and LH after treatment with 0.005% estriol vaginal gel will be studied in each woman using Wilcoxon-signed rank test. The variation of levels between two arms will be analyzed using a non-parametric test (Mann-Whitney-Wilcoxon test) or an ANCOVA (if it must be adjusted by initial value).

Adverse events data and serious adverse events will be reported in frequency tables. The safety analysis will be performed in the population that has received at least one dose of the drug. AEs

will be graded according to MedDRA. Adverse events between groups will be compared using the chi-square tests of independence (Fisher's Exact test in the case of observing frequencies <5%).

Symptoms of Vaginal atrophy will be evaluated with a codification from 0 to 3.

The variations in the intensities for each one of the symptoms and signs of the vaginal atrophy, after 3 and 12 weeks, in each treatment arm, will be compared using a non-parametric test (Mann-Whitney-Wilcoxon). Descriptive statistics will be calculated for all variables studied. A global symptoms score will be evaluated; it will be calculated by summing the intensities of all the three symptoms of vaginal atrophy in a certain time point.

Sexual function measured by the FSFI [5] scale. It will be used an algorithm for determining domain scores and a composite full-scale score.

Time between previous study medication administration dose time and blood sample (serum and plasma) collection time will be summarized (mean, median, range,..) for each of visits and will be compared using a non-parametric test (Wilcoxon-signed rank test).

Primary and secondary objectives will be analyzed in the same way, without those patients who may cause baseline differences, in a sensitivity analysis to evaluate robustness of the conclusions.

The significance level of all statistical tests is established at 0.05.

9.2.1 Patient Disposition

A detailed description of patient disposition will be provided. It could include the following:

- summary of patients entered and by site
- total number of patients entered
- total number of patients enrolled
- summary of reasons for patients entered, but not enrolled
- total number of patients treated
- summary of reasons for patients enrolled, but not treated.

A detailed summary of reasons for patient discontinuation from study treatment will be provided.

A summary of all identified important protocol violations will be provided.

9.2.2 Patient Characteristics

Patient characteristics could include a summary of the following:

- patients demographics
- time of menopause
- baseline disease characteristics
- preexisting conditions/secondary conditions
- prior therapy

Other patient characteristics will be summarized as deemed appropriate.

The categorical variables will be summarized by contingency tables. The continuous variables will be described by the number of patients with valid values (n), mean, standard deviation, median and range in both arms.

The effect of the treatment on continuous variables will be assessed by Mann-Whitney-Wilcoxon test. Categorical variables will be analyzed by chi-square or Fisher's exact test.

9.2.3 Concomitant Therapy

A summary of concomitant therapies will be generated in the safety population.

9.2.4 Treatment Compliance

The percentage of complying patients will be calculated at weeks 3, 8 and 12, defined as those with a compliance percentage of 80-110%, in the group of patients treated with 0.005% Estriol vaginal gel and in the placebo group. The overall compliance of the study drug will be calculated from these partial calculations and in the population that has received at least one dose of the drug. It will be summarized the patients and reasons of those patients which did not complete the 80% of medication of the study.

The estimate of percentage compliance will be given by:

$$Percent\ Compliance\ = \frac{Actual\ dose\ administered\ per\ week}{Dose\ expected\ to\ be\ administered\ per\ week}\ x\ 100$$

According to the protocol:

Weeks 1-3 single daily application 1 g of gel (7g/week).

Weeks 4-12 twice weekly administration 1 g of gel (2g/week).

Dose expected to be administered: (7gx3w) + (2gx9w) = 39g in 12 weeks

9.2.5 Hormone levels

All these analysis will be based on the ITT population.

The primary endpoint and one of the secondary endpoints are to evaluate the variation from baseline of FSH, LH, estriol, estradiol and estrone blood levels in patients treated with 0.005% estriol vaginal gel.

Two determinations of FSH and LH before treatment with 0.005% estriol vaginal gel will be studied in each woman in order to study the intra-individual variation. Screening and baseline determinations will be used for this purpose. This intra-individual variation will be compared to the variation between baseline and the values obtained at every study visits to study the FSH and LH variation caused by treatment.

Changes in hormone levels will be studied in each woman along the treatment. Also, changes in hormone levels will be compared by arm.

9.2.6 Main Study Analysis

The primary endpoint, the variation of FSH between week 12 and baseline and also comparing by arm will be analyzed using a non-parametric test (Mann-Whitney-Wilcoxon test).

An ANCOVA analysis will be used to corroborate the results obtained for the primary endpoint. The ANCOVA analysis will compare the variation of FSH 12 weeks adjusted by FSH basal between the two arms. This analysis will be performed if the relation between FSH 12 and FSH basal is lineal.

These analyses will be evaluated in the ITT population, and also in Per-Protocol population to confirm the results.

9.2.7 Efficacy analyses

9.2.7.1 Signs and symptoms of vaginal atrophy

The symptoms evaluated will be the following:

- Vaginal dryness
- Pruritus or itching
- Dyspareunia

Each symptom will be scored in a numeric scale from 0 to 3, as shown below:

- 0 Absence. The symptom is not present
- 1 The symptom is of mild intensity, without interfering in the patient's activity
- 2 The symptom is of moderate intensity, causing obvious discomfort to the patient
- 3 The symptom is stated as very irritating and severe in intensity

A Global Symptoms Score will be evaluated. It will be calculated by summing the intensities of all the three symptoms of vaginal atrophy in a certain time point, thus ranging between 0 and 9. The variation (difference between the sum of the intensity in week 3 or 12 minus the sum of the intensity in baseline) in the Global Symptom Score vs baseline will be assessed at weeks 3 and 12.

Two different versions of this analysis will be done taking into consideration if the patients have had previous sexual intercourse:

- Version 1: dyspareunia will not be score for patients that did not have previous sexual intercourse, and therefore it will not be taken into consideration for the global score and the statistical analysis.
- Version 2: dyspareunia will be score as 0 for patients that did not have previous sexual intercourse.

Version 1 will be the one used for the evaluation and interpretation of the study results.

The signs evaluated will be the following:

- Vaginal mucosa with flattening of folds or thinning

- Dryness of the mucosa
- Fragility of the mucosa

Each one of these signs will be scored by the investigator on a numerical scale in accordance with their presence and degree of severity as follows:

- 0 Absence. The sign is not present.
- 1 The sign is present and is considered a mild alteration
- 2 The sign is present and is considered a moderate alteration
- 3 The sign is present and is considered a severe alteration

The variations in the intensities for each one of the symptoms and signs of the vaginal atrophy, after 3 and 12 weeks, in each treatment arm, will be compared using the non-parametric test Mann-Whitney-Wilcoxon.

9.2.7.2 Vaginal Maturation value and pH

The averages of the differences between the baseline and the final MV and pH of the 0.005% Estriol vaginal gel and placebo groups will be shared by a non-parametric test, to determine the possible superiority of the 0.005% Estriol vaginal gel treatment compared with placebo administration.

These same tests will be used to analyse the change in the MV and pH after the initial observation period of 3 weeks.

All of the above secondary analyses will be conducted at a two-sided 0.05 level of significance.

9.2.8 Safety Analyses

The safety analysis will be evaluated in the safety population.

The toxicity and tolerability of the study drug will be evaluated in the safety population. Safety analyses will include summaries of the incidence of adverse events by MedDRA that occur during the study treatment period or within 30 days (+/- 5 days) of the last dose of study treatment, regardless of causality and according to the relationship to study drug as assessed by the investigator. Additionally, the following safety-related outcomes will be summarized:

- study treatment discontinuations due to adverse events.
- deaths
- SAEs
- hospitalizations and transfusions
- use of key concomitant medications or growth factors.

Analyses for data with discrete dates, for example, deaths, transfusions, and concomitant medications, will be done through 30 days after each patient's last dose of study treatment. Adverse events will also be analyzed in this timeframe; that is, if an event starts within 30 days of discontinuation from study treatment, but after 30 days after the last dose of study treatment, it will not be included.

Adverse events data and serious adverse events will be presented in frequency tables by grade. Hematological and clinical biochemistry toxicities will be assessed from laboratory test parameters. The safety analysis will be performed in the safety population.

Adverse events data and serious adverse events will be reported in frequency tables (overall and by grades). Adverse events will be compared using the chi-square tests (Fisher's Exact test in the case of observing frequencies <5%).

9.2.9 Other Analyses: Patient Reported Outcomes: FSFI

Sexual function measured by the Female Sexual Function Index (FSFI) scale, a 19-item questionnaire, has been developed as a brief, multidimensional self-report instrument for assessing the key dimensions of sexual function in women. It will be used an algorithm for determining domain scores and a composite full-scale score. The individual domain scores and full scale (overall) score of the FSFI can be derived from the computational formula outlined in the table below. Individual domain scores are obtained by adding the scores of the individual items that comprise the domain and multiplying the sum by the domain factor. The full scale score is obtained by adding the six domain scores. It should be noted that within the individual domains, a domain score of zero indicates that no sexual activity was reported during the previous evaluation.

Domain	Questions	Score Range	Factor	Minimum Score	Maximum Score	Score
Desire	1, 2	1 – 5	0.6	1.2	6.0	
Arousal	3, 4, 5, 6	0 – 5	0.3	0	6.0	
Lubrication	7, 8, 9, 10	0 – 5	0.3	0	6.0	
Orgasm	11, 12, 13	0 – 5	0.4	0	6.0	
Satisfaction	14, 15, 16	0 (or 1) – 5*	0.4	0.8	6.0	
Pain	17, 18, 19	0 – 5	0.4	0	6.0	
		Full Scale Score Range		2.0	36.0 Total	

*Range for item 14 = 0–5; range for items 15 and 16 = 1–5

A score ≤ 26.55 is classified as FSD (Female Sexual Dysfunction)

10. TABLES AND FIGURES

Table/Figure No.		Title of Table/Figure
1.		*Recruitment of Site*
2.		*Consort Flowchart*
	2.1	*Protocol deviations*
	2.2	*Analysis Populations*
3.		*Characteristics of patients*
4.		*Treatment*
5.		*Safety analysis*
6.		*End of treatment*
7.		*Efficacy Analysis*
8.		*Signs and symptoms of vaginal atrophy analysis*
9.		*Vaginal Maturation value and pH analysis*
10.		*Patient Reported Outcomes: FSFI analysis*

The numbers of tables and figures do not have to match exactly with those of the statistical report.

11. BIBLIOGRAPHY AND REFERENCES

[1] GEICAM, "Protocol ITFE-2026-C10".

[2] "ICH E9, Statistical Principles for Clinical Trials.".

[3] "ICH E3, Structure and Content of Clinical Study Reports".

[4] G. Pfeiler, C. Glatz, R. Königsberg, T. Geisendorfer, A. Fink-Retter, E. Kubista, C. F. Singer and M. Seifert, "Vaginal estriol to overcome side-effects of aromatase inhibitors in breast cancer patients," *Climateric,* pp. 339-344, 2011.

[5] R. Rosen, C. Brown and J. Heiman, "The Female Sexual Function Index (FSFI): A Multidimensional Self-Report Instrument for the Assessment of Female Sexual Function," *Journal of Sex & Marital Therapy,* vol. 26, no. 191-208, pp. 191-208, 2000.

[6] A. C. Hope, «A Simplified Monte Carlo Significance Test Procedure,» *Journal of the Royal Statistical Society,* vol. 30, nº 3, pp. 582-598, 1968.

[7] H. &. N. N. Dette, «Nonparametric analysis of covariance,» *The Annals of Statistics,* vol. 29, nº 5, p. 1361–1400, 2001.

[8] S. A. Y. D. M.G. Akritas, «Nonparametric models and methods for nonlinear analysis of covariance,» *Biometrika,* vol. 87, nº 3, p. 507–526, 2000.