KARISMA II

Statistical Analysis Plan

October 25, 2016

Approvals

Author	Martin Eklund, statistician	
Approved by	Per Hall, PI	
	Kamila Czene, co-PI	

Revision history

Version	**Date**	**Reason for change**
0.3	October 17, 2016	First draft
0.4	October 25, 2016	Split of primary analysis

Contents

1 Preface

This Statistical Analysis Plan (SAP) describes the planned analyses for *Karisma II: A randomized, double blinded, six-armed placebo controlled study to investigate optimal dose of tamoxifen with the most favourable side effect spectra and with mammographic density reduction non-inferior to that of 20 mg tamoxifen*, EudraCT number 2016-000882-22.

The planned analyses identified in this SAP will be included in future manuscripts. Exploratory analyses not necessarily identified in this SAP may be performed to support planned analyses. Any post-hoc exploratory or unplanned analyses not specified in this SAP will be identified as such in manuscripts for publication.

This SAP was written by a statistician and investigators who were blinded to treatment allocation and treatment-related study results and will remain so until the central database is locked and the final data are extracted for analysis. To ensure blinding, treatment allocations are stored in a separate location accessible only by an un-blinded statistician.

2 Abbreviations

Abbreviation	**Explanation**
SAP	Statistical analysis plan
ITT	Intention to treat
PP	Per protocol
MP	Multiple imputation

3 Background

For more than 40 years, 20 mg of tamoxifen has been used in breast cancer patients to reduce the risk of recurrence. It has also been shown that 20 mg of tamoxifen reduces the risk of breast cancer (by 40-60%) in healthy women but that the treatment comes with side effects. The side effects are probably one reason to that tamoxifen is not used in the preventive setting despite the substantial risk reducing effect. The aim of the Karisma II study is to identify the lowest tamoxifen dose that could be used for primary preventive purposes, with a hypothesis that a lower dose will reduce the side effects.

Mammographic density is the white part of the mammogram and consist of glandular and connective tissue. The black, non-radiolucent part is fat.

The more glandular and more connective tissues, the whiter the breast, and the higher the risk of breast cancer [1]. Factors that influence risk of breast cancer are known to influence density. Hormone replacement therapy increases density and risk of breast cancer, while tamoxifen has the opposite effect. The molecular basis for the tamoxifen-induced changes in breast density is sparsely understood.

There are three studies that used mammographic density change during one year as a proxy for tamoxifen effect [2–4]. Two projects studied the association between mammographic density change in patients and prognosis and one study looked at density change among perfectly healthy woman and risk of breast cancer. All three studies came to the same result, approximately half of the treated participants respond to the therapy by a significant reduction in mammographic density after one year. Those that respond with a decreased mammographic density had a lower risk of breast cancer recurrence and a lower risk of breast cancer.

The conclusion is that change in mammographic density is a very good marker of Tamoxifen therapy response.

4 Design

This is a dose determination study aiming to identify the optimal tamoxifen dose for reducing the risk of breast cancer. Instead of following a very large number of women, for many years, treated with different doses of tamoxifen and see if they develop breast cancer, we will include 1440 healthy women participating in the mammographic screening program at two Mammography units (Unilabs mammography, Lund and Södersjukhuset Breast Centre, Stockholm) and measure their change in mammographic density. We will test if 1 mg, 2.5 mg, 5 mg and 10 mg reduce the mammographic density to the same extent as 20 mg. In summary Karisma II will be a randomized, double-blinded, six-armed placebo controlled study to identify the dose of tamoxifen with the most favourable side effect spectra and with a density reduction non-inferior to 20 mg tamoxifen.

Secondary aims of the Karisma II trial are to investigate changes in tissue markers by comparing breast tissue biopsies sampled prior to and after six months of treatment with tamoxifen or placebo, as well as relate levels of tamoxifen metabolites, proteins, lipids and hormones in blood and changes in breast tissue to tamoxifen doses. Moreover, we will study genetic polymorphism in germline DNA and their association to the other endpoints.

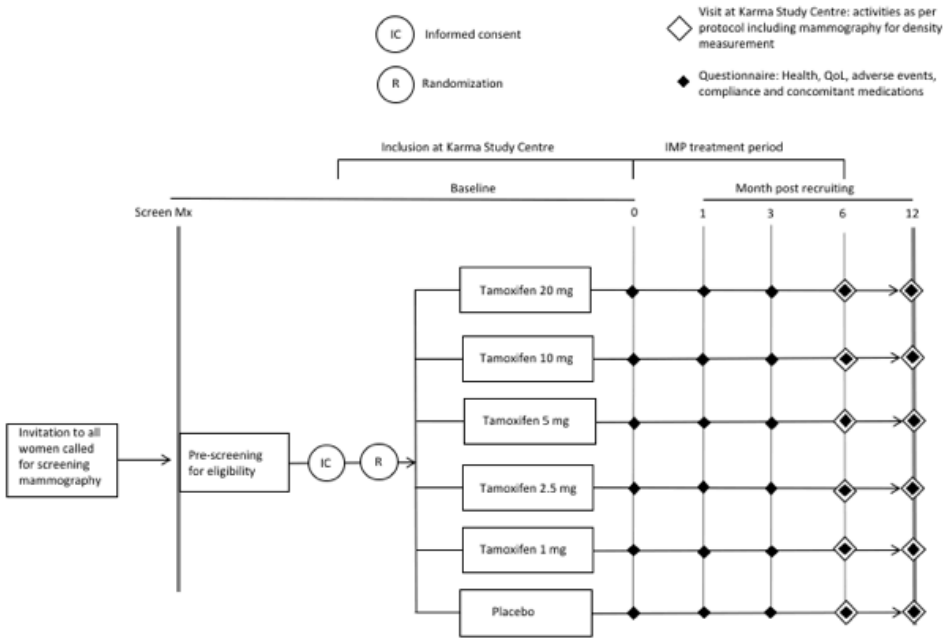

Figure 1: Overall study design.

4.1 Study population

The study will include healthy women that participate in the mammographic screening program at Södersjukhuset, Stockholm and at Skånes Universitetssjukhus, Lund, and have a negative screening mammogram. We will not include breast cancer patients as they receive a variety of drugs that might influence mammographic density. We are excluding women with no measurable density (< XXX *** according to the Stratus breast density measurement method *** this is 4.5% or more by Volpara according to the protocol (p. 17), is this still the way it should be done?). This is because measurable density at study start is a prerequisite for being able to see reduction of mammographic density over time. We are not selecting women based on risk of breast cancer since we currently lack a reliable risk prediction model for Swedish standards.

4.2 Study period

Karisma I was a proof of concept study preceding Karisma II. The main reason we conducted the Karisma I study was to investigate time to density change. Radiologists tell us that the time to density change is most probably much shorter than one year, probably only a couple of months. However, the issue of time to density change during tamoxifen therapy has never been addressed.

The results from Karisma I show that, regardless of dose, density decreases within 6 months by about 50% for one quarter of the participants, about 20% for another quarter and not at all for the other participants. This result implicates that tamoxifen intake during 6 months gives a measurable density change of responders; therefore, the treatment period in Karisma II are set to 6 months. From a clinical point of view and for coming study (Karisma III) we will investigate the state of density change and decline of side effects 6 months after tamoxifen cessation in relation to doses.

5 Endpoints

5.1 Primary objective

Identify the minimal dose of tamoxifen that is non-inferior in its ability to reduce mammographic density compared with 20 mg tamoxifen.

5.2 Primary endpoint

Change in mammographic density. In particular, we will test for noninferiority in the proportion of women in the intervention arms (placebo, 1 mg, 2.5 mg, 5 mg, 10 mg) who have a density reduction as great as or greater (after 6 months) than the median density reduction in the 20 mg arm (see Section 6.3).

5.3 Secondary objectives

Assess the drop-out level and the level of side effects in the intervention arms compared to the 20 mg arm.

5.4 Secondary endpoint

Drop-out rate. We will test for differences in the proportion of drop-outs after 6 months in the intervention arms compared to the 20 mg arm.

5.5 Tertiary objectives

Relate levels of tamoxifen metabolites, proteins, lipids and hormones in blood and changes in breast tissue [*** according to what measure] to tamoxifen doses. Measure genetic polymorphism in germline DNA and relate it to the other secondary objectives. Measure mammographic density and levels of side effects 6 months after tamoxifen cessation in relation to tamoxifen dose. Study associations between dose of tamoxifen, polymorphisms in CYP2D6 and blood levels of endoxifen. Quantify the level of side effects (according to the FACT-ES instrument) and compliance of medicine intake (Morisky scale).

5.6 Tertiary endpoints

Levels of tamoxifen metabolites (e.g. endoxifen), proteins [*** which?], lipids [*** which?] and hormones [*** which?] in blood. Breast tissue changes. Genetic polymorphism in germline DNA, in particular CUP2D6. Mammographic density and levels of side effects 6 months after tamoxifen cessation. Side effects (according to the FACT-ES instrument) and compliance of medicine intake (Morisky scale).

6 Statistical methods

All analyses will be performed after the study is completed and the database is released. All statistics, including tables, figures and listings, will be performed using SAS, Version 9.3, or R, Version 3.x.

6.1 Populations

The intention to treat (ITT) population consists of all women who were randomized and had eligible measure at baseline and at least one eligible measure post baseline. The Per Protocol (PP) Population consists of all women in the ITT population who remained in the study for at least 6 months (the time period during which Tamoxifen is administered) in the study and who had no significant protocol violations (compliance of study drug intake is defined as at least 80% of the prescribed numbers of tablets should be used during time of treatment). The analyses will be preformed on both ITT and PP, however the primary analyses will be carried out on an ITT basis. ITT analysis will likely bias the results toward equivalence for the primary efficacy endpoint compared to if all participants completed the

study (assuming that non-compliance is dose dependant). The per-protocol analysis, on the other hand, excludes data from patients with major protocol violations and may bias the results in either direction. We will therefore analyse the primary endpoint using both intention-to-treat and per-protocol approaches.

Since we expect around 30% dropout rate, we will also create a dataset where missing data is imputed; the multiple imputation (MP) population (see Section 6.9 for details).

6.2 Demographics and baseline data

All data will be presented using descriptive statistics. Results will be presented in total, by treatment group, age and menopausal status. Continuous variables will be summarized using number of women, mean, standard deviation, median, minimum and maximum. Categorical variables will be summarized using the number and percentage of women.

6.3 Primary analysis: Proportion of responders (efficacy)

The results of both papers [2] and [3] showed that a threshold value in mammographic density decrease needs to be exceeded in order for a woman to experience the beneficial effect of tamoxifen. This threshold value was different in [2] compared to [3]. A 10% or greater reduction was needed in [2] whereas a 20% reduction was needed in [3], a difference most likely explained by different methods of measuring mammographic density. However, the proportion of women exceeding this threshold was exactly 48% in both studies, implying that approximately 50% of the women experience the beneficial effect of tamoxifen irrespective of the way mammographic density is measured.

In the Karma II study, we will use the Stratus method for measuring breast density, which was not used in either [2] or [3]. With support of the results in [2] and [3], we define the response threshold to be the median decrease in mammographic density of women in the 20 mg tamoxifen arm (this definition ensures that the same proportion of women exceeds the threshold as observed in previous studies). Further, we define a responder to be a woman whose mammographic density decrease between baseline and 6 months exceeds the response threshold.

The primary efficacy endpoint is the proportion of responders (women with a reduction of mammographic density, from baseline to 6 months). We will test for non-inferiority with respect to the primary efficacy endpoint for

groups of women treated with placebo, 1, 2.5, 5, and 10 mg compared to the group of women treated with 20 mg tamoxifen. We will use Fisher's exact test with an p-value of less than 0.05 (corrected for multiple testing, see Section 6.7) as a threshold to declare statistical significance. The non-inferiority margin is defined to be 16.7 percentage points; that is, we define non-inferiority to mean that the fraction of responders is not less than one third of the treated individuals (50% minus 16.7% = 33.3%). The null hypothesis is thus that the proportion of responders in women treated with placebo, 1, 2.5, 5, and 10 mg is 16.7 percentage points less than those women treated with 20 mg tamoxifen (i.e. the fraction of responders is less than one third). For rejection of the null hypothesis, the lower end of the confidence interval needs to by higher than -16.7%. For this to hold, the point estimate of an intervention arm can not be more than 5 percentage points lower than the percent responders in the 20 mg arm.

6.4 Secondary analysis: Side effects (safety)

We will assess the proportion of drop-outs at six months and test for differences in the proportion of drop-outs in the intervention arms compared with the 20 mg arm. We will use Fisher's exact test with an p-value of less than 0.05 (corrected for multiple testing, see Section 6.7) as a threshold to declare statistical significance.

6.5 Tertiary analyses

We will perform two types of tertiary analyses: (i) sensitivity analysis on the primary and secondary endpoints using different methods than used in the primary and secondary analyses in order to assess sensitivity to choice of analyses pathway, assumptions, and missing data; and (ii) on tertiary endpoints.

6.5.1 Sensitivity analyses of the primary endpoint

1. We will perform sensitivity analyses to the primary analyses where the threshold for being defined as a responder is varied ± 10 percentage points, to assess whether the overall conclusion of the trial is sensitive to the within study definition of what a *responder* is. Similarly, we will perform a tipping point analysis, where the the threshold for being defined as a responder is changed until the result of the trial changes (either from a nonsignificant result to a significant result or vice versa).

2. Using absolute change in density instead of relative change to define a responder.

3. We will analyze data from all arms in a regression model (as opposed to one arm against another arm). I.e., responder status (a binary variable) will be used as the response variable in a logistic regression model, where dose allocation is the independent variable.

4. We will impute missing data (as described in Section 6.9) and re-perform the primary analysis using imputed datasets (in a multiple imputation protocol).

5. We will analyse the results using the recorded continuous endpoint (density reduction) instead of discretising the reduction into either *responder* or *non-responder*. the We will use both relative and absolute density reduction is endpoints in this analysis.

6.5.2 Sensitivity analyses of the primary and secondary endpoints simultaneously

We will perform an analysis where the distance to the diagonal line in a plot of proportion responders vs proportion drop-outs by study arm is quantified (see Figure 2 for a schematic mock figure). Putatively, the arm with the greatest distance to the diagonal line gives the best tradeoff between efficacy and drop-out level. Confidence intervals of the distance will be computed using the bootstrap.

6.5.3 Tertiary endpoints

Tertiary outcomes will be analysed with statistical methods appropriate for the specific outcome. A large number of tertiary endpoints are being recorded within the trial, e.g. levels of tamoxifen metabolites, proteins, lipids and hormones in blood, breast tissue changes, genetic polymorphism in germline DNA, mammographic density and levels of side effects 6 months after tamoxifen cessation. Since this the first study of its type, the associations of nearly every possible combination of primary, secondary, and tertiary endpoints are of interest. We therefore expect a number of tertiary analyses to be performed. In the list below, we briefly specify some obvious examples of such analyses:

1. The association of dose of tamoxifen to blood concentration of tamoxifen metabolites (e.g. endoxifen) by CYP2D6 polymorphism status (a woman can be classified as an excellent, intermediate, or poor

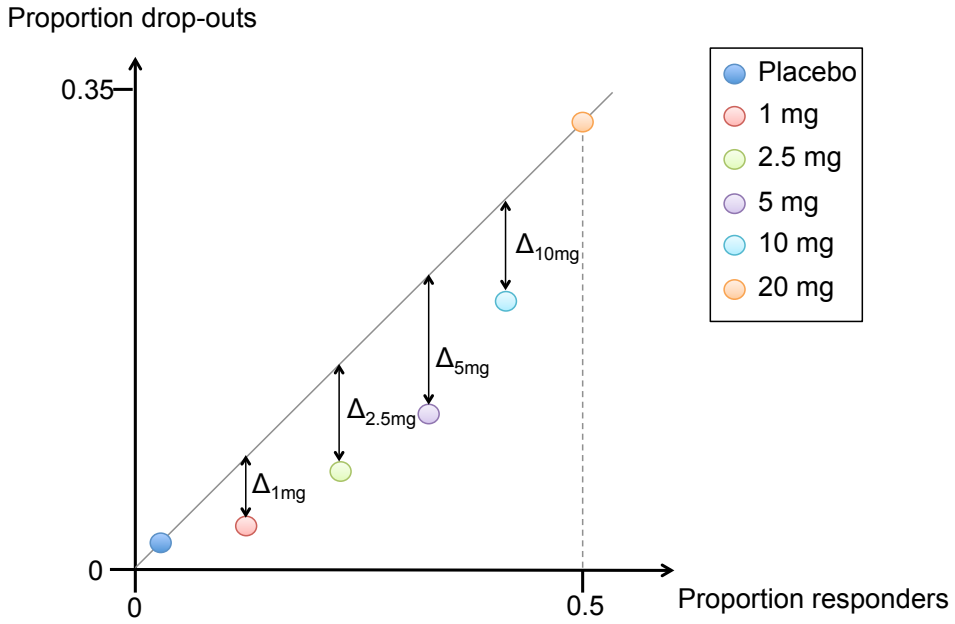

Figure 2: Schematic picture of what a plot of proportion responders vs proportion drop-outs by study arm may look like. The arm with the greatest distance to the diagonal line gives the best tradeoff between efficacy and drop-out level.

metaboliser of tamoxifen based on polymorphism patterns in CYP2D6 [5]) will be analysed in a linear regression model with tamoxifen dose, CYP2D6 polymorphism status, and age as independent variables.

2. The association of plasma concentrations of endoxifen with responder status will be analysed with logistic regression both within treatment groups and among treatments groups (treatment group will be handled as a random effect in the latter case).

6.6 Sample size

Since we have five intervention arms and since we do the primary analysis on a "per arm basis", we have five chances to show noninferiority of the primary endpoint and we therefore need to correct for multiple testing. Standard methods for correcting for multiple testing (according to the methods by Bonferroni or Holm) are too conservative in this context, since they

rely on an assumption of independence between the tests whereas the tests here are clearly correlated. We will instead use the resampling based step-down method suggested by Romano, Shaikh, and Wolf [6], which seamlessly handles correlated tests (see also Section 6.7). The correlation structure of the tests is unknown and we therefore assume that the multiplicativity leads to that the α decreases from 0.05 to 0.03 in order to declare a significant results (correction by Bonferroni would lead to an α of 0.01).

With 168 participants in each study arm, the lower limit of the observed one-sided 98.5% confidence limit for the difference will be expected to exceed the non-inferiority margin 16.7 percentage points with a 80% power, assuming that the expected difference in responders between two arms is 0 and the proportion responders in the 20 mg tamoxifen group is 50%. With an estimated 30% drop-out rate, 240 patients in each group need to be randomized.

The trial is powered for the primary analysis. However, we here also briefly investigate the power for showing *both* noninferiority with respect to the primary endpoint and superiority with respect to the secondary endpoint. With 240 participants randomized to each study arm, we have 79% power to show both noninferiority with respect to the primary endpoint and superiority with respect to the secondary endpoint assuming that the proportion of drop outs is reduced by two thirds (from 30% to 10%) and 67% power if the proportion of drop outs is reduced by 50% (from 30% to 15%). (For these calculations, we use the same assumptions about the correlation of the tests as for the analysis of primary endpoint.)

The computations are based on the following formula:

$$n = f(\alpha, \beta)\frac{\pi_s(100 - \pi_s) + \pi_e(100 - \pi_e)}{(\pi_s - \pi_e - d)^2},$$

where π_s and π_e are the true percent 'success' in the standard (20mg) and experimental treatment group respectively, and

$$f(\alpha, \beta) = \left[\Phi^{-1}(\alpha) + \Phi^{-1}(\beta)\right]^2,$$

and Φ^{-1} is the cumulative distribution function of a standardised normal distribution.

6.7 Adjustment for multiplicity, confounders, and heterogeneity

Since we have five treatment arms and therefore five comparisons, we will need to adjust for multiple testing. The treatment arms are correlated,

since they are defined by different doses of the same drug. Thus, standard adjustment for multiple testing using the methods by Bonferroni or Holm are too conservative. Instead, we will use the bootstrap based method suggested by Romano, Shaikh, and Wolf [6], which seamlessly handles correlated tests. The correlation structure of the tests is unknown and we therefore assume that the multiplicativity leads to that the α decreases from 0.05 to 0.03 in order to declare a significant results (correction by Bonferroni would lead to an α of 0.01).

We will perform subgroup analyses, where the subgroups are CYP2D6 polymorphism status (excellent, intermediate, poor metaboliser [5]). CYP2D6 polymorphism status will be handled as a categorical fixed effect variable. Other subgroups will likely be studied in non-prespecified analyses.

6.8 Interim analysis

We will three months into the study perform an interim analysis on overall dropout level. If the proportion of drop out at three months, we will – at best – have 73% power for our primary analysis (under the same assumptions as described in Section 6.6, apart from the assumption about the proportion of drop outs). In this case, we will increase the number of women the study needs to accrue according to:

$$n_{tot} = 168 \times 6/(1 - d'),$$

where d' is the overall proportion of drop outs at three months.

6.9 Handling of missing data

The primary analysis will be performed on an ITT basis using the last-observation-carried-forward method. This analysis makes sense from an effectiveness (*de facto*) point of view, however it will likely bias the results towards equivalence for the primary efficacy endpoint (assuming that non-compliance is dose dependant) from an efficacy (*de jure*) point of view (i.e. compared to if no women dropped out from the study). We will therefore perform sensitivity analyses (Section 6.5) where we impute mammographic density 6 months into the study for women who dropped out. Specifically, we will use a multiple imputation protocol where mammographic density at 6 months is imputed using a regression model with density at the time points recorded within the study, age, HRT status, CYP2D6 polymorphisms, and endoxifen levels (at the available time points). The exact choice of multiple imputation model will be done using the study data after database lock using

cross-validation. We will also perform tipping point analyses with respect to the missing data by exploring how extreme the imputation of the missing data will need to be to change the overall conclusion of the trial (see also Section 6.5.1).

References

[1] Norman F Boyd, Helen Guo, Lisa J Martin, Limei Sun, Jennifer Stone, Eve Fishell, Roberta A Jong, Greg Hislop, Anna Chiarelli, Salomon Minkin, and Martin J Yaffe. Mammographic density and the risk and detection of breast cancer. *N Engl J Med*, 356(3):227–36, Jan 2007. doi: 10.1056/NEJMoa062790.

[2] Jack Cuzick, Jane Warwick, Elizabeth Pinney, Stephen W Duffy, Simon Cawthorn, Anthony Howell, John F Forbes, and Ruth M L Warren. Tamoxifen-induced reduction in mammographic density and breast cancer risk reduction: a nested case-control study. *J Natl Cancer Inst*, 103 (9):744–52, May 2011. doi: 10.1093/jnci/djr079.

[3] Jingmei Li, Keith Humphreys, Louise Eriksson, Gustaf Edgren, Kamila Czene, and Per Hall. Mammographic density reduction is a prognostic marker of response to adjuvant tamoxifen therapy in postmenopausal patients with breast cancer. *J Clin Oncol*, 31(18):2249–56, Jun 2013. doi: 10.1200/JCO.2012.44.5015.

[4] Sarah J Nyante, Mark E Sherman, Ruth M Pfeiffer, Amy Berrington de Gonzalez, Louise A Brinton, Erin J Aiello Bowles, Robert N Hoover, Andrew Glass, and Gretchen L Gierach. Prognostic significance of mammographic density change after initiation of tamoxifen for er-positive breast cancer. *J Natl Cancer Inst*, 107(3), Mar 2015. doi: 10.1093/jnci/dju425.

[5] Jingmei Li, Kamila Czene, Hiltrud Brauch, Werner Schroth, Pilar Saladores, Yi Li, Keith Humphreys, and Per Hall. Association of cyp2d6 metabolizer status with mammographic density change in response to tamoxifen treatment. *Breast Cancer Res*, 15(5):R93, 2013. doi: 10.1186/bcr3495.

[6] Joseph P. Romano, Azeem M. Shaikh, and Michael Wolf. Control of the false discovery rate under dependence using the bootstrap and subsampling. *TEST*, 17(3):417, 2008. ISSN 1863-8260. doi: 10.1007/s11749-008-0126-6. URL `http://dx.doi.org/10.1007/s11749-008-0126-6`.