Mathematical Statistics
Stockholm University

Endpoints in Clinical Trials Investigating the Use of Hormone Replacement Therapy in Menopausal Women

Marie Göthberg

Examensarbete 2006:18

Postal address:

Mathematical Statistics
Dept. of Mathematics
Stockholm University
SE-106 91 Stockholm
Sweden

Internet:

http://www.math.su.se/matstat

Mathematical Statistics
Stockholm University
Examensarbete **2006:18**,
http://www.math.su.se/matstat

Endpoints in Clinical Trials Investigating the Use of Hormone Replacement Therapy in Menopausal Women

Marie Göthberg*

October 2006

Abstract

This essay aims to describe endpoints used in clinical trials to assess safety of hormone therapy prescribed to women with postmenopausal symptoms. A number of endpoints for the bleeding profile are presented, discussed and analysed using different statistical models, with data from a clinical trial used to illustrate the statistical methods. Several suggestions for how to present and illustrate this type of data are given. The currently drafted guidelines on the subject from the European and U.S. authorities are discussed.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: perra.marie@telia.com. Supervisor: Juni Palmgren.

Contents

List of Tables

List of Figures

Introduction

This essay aims to describe endpoints used in clinical trials to assess safety of hormone therapy prescribed to women with postmenopausal symptoms. The central part describes and compares different statistical methods for how the bleeding profile for two different treatments can be compared and analysed. Measures of efficacy are briefly touched upon.

The first chapter includes the basic information about the therapeutic indication, about different hormone therapies and the potential breast cancer risk associated with this therapy.

The most commonly used safety endpoint is the bleeding profile. This endpoint is described and discussed in the second chapter. From the bleeding profile several other relevant endpoints are defined. Some issues regarding how data is collected and potential problems with regard to these endpoints are also discussed here.

Chapter three is the central part of this essay and here active treatment is compared to placebo with regard to the bleeding profile. The bleeding profiles are compared in terms of proportions, odds ratios, hazard ratios and rate ratios using different statistical methods and models. Data from an older trial are used to illustrate the different statistical methods. In these data the age-distributions for the two treatments differ, with younger women in the active treatment group. This imbalance need to be adjusted for since the rate for bleeding decreases with increasing age.

In chapter four the currently drafted guidelines on hormone therapy from the European and the American authorities are presented and discussed, in particular with regard to the bleeding profile. A test strategy for multiple efficacy endpoints is suggested.

The fifth and the last chapter contain a summary and a short discussion.

Acknowledgement

I would like to express my sincere gratitude to my supervisor Juni Palmgren for helpful suggestions, stimulating discussions and encouragement in the writing of this 10 credit essay.

1 Hormone Replacement Therapy

1.1 Menopause

A woman enters the menopause when the menstrual bleeding has stopped and she is no longer of childbearing potential. When the menopause is initiated the production of estrogen is considerably decreased and follicle-stimulating hormone (FSH) is considerably increased. At the same time many women start to experience symptoms such as: hot flushes (hot flashes), palpitations, depression, nervousness, sleeping difficulties, headaches and muscle and joint-pain [1]. Hot flushes are commonly referred to as vasomotor symptoms and are experienced by more than 75% of the women in the menopause. The vasomotor symptoms are usually resolved within 2 years but for 20% of the women it can last for more than 5 years [2]. A hot flush can be anything from a sensation of heat to an intense feeling of pressure over the breast, the neck and the head which is accompanied by profuse flushing and sweating.

1.2 Hormone Replacement Therapy (HRT)

Although the mechanisms behind hot flushes are still not known, hormone replacement therapy (HRT) has been used for more than 60 years to effectively treat these and other menopausal symptoms. The use of HRT grew dramatically in the seventies after it had been shown in several clinical trials that treatment with estrogen maintained the bone mineral density. This new HRT indication made it accepted to prescribe HRT to women with an increased risk of osteoporosis as a prophylaxis and therefore lifelong treatment. The general attitude towards HRT was positive and it was also believed, until a few years ago, that HRT had a preventive effect on heart disease, heart attacks and stroke [3].

HRT are available in several different administrations such as patches, implants, oral tables and vaginal tablets. The HRT can have a local or a systemic effect. The preparations are available in a large variety of combinations of different hormones, mainly including combinations of estrogen and progestogen. The reason for adding progestogen is that it normalises the risk for obtaining endometrial cancer, which was shown to be increased for products containing only estrogen in the seventies. Estrogen alone products should therefore only be used by women without an uterus.

Several systemic HRT regimens are available for non-hysterectomized postmenopausal women, including sequential and continuous regimens. A sequential HRT is characterized by a continuous administration of estrogen but where a progestogen is added for 10-14 days during the cycle. This type of HRT regimen causes a monthly menstrual-like withdrawal bleeding in the

great majority of women.

A continuous HRT regimen is characterized by daily continuous administration of progestogen in combination with estrogen. This type of HRT regimen does not cause any monthly withdrawal bleeding. However, during continuous HRT occasional bleedings have been reported during the initial 6 months of treatment. These bleedings seem to be more frequent for continuous HRT regimens with higher doses. Minimization of occasional bleedings associated with continuous HRT may be an important factor for treatment adherence.

The HRT products that have been used to treat menopausal symptoms and osteoporosis are mainly systemic and continuous or sequential treatments. Most of the oral products available today are taken daily and have an estrogen dose of between 0.5 and 2.0 mg.

The name, Hormone Replacement Therapy, is suggesting that the therapy normalises the hormone levels. However, since the estrogen production by nature is ceased for a postmenopausal woman it would be more appropriate to refer to the therapy as "only" Hormone Therapy (HT).

1.3 Breast Cancer

Breast cancer is the most common cancer in women in the world today [4]. It is also the type of cancer that causes the most deaths. According to official Swedish cancer statistics for 2004 (Socialstyrelsen accessible via the web [5]) a total of 6 925 new cases of breast cancer were diagnosed in women. The majority of these cases (81%) were diagnosed in women likely to be postmenopausal, i.e. older than 50 years. The incidence was 142 cases per 100 000 women, which corresponds to a 34% increase in the incidence for the last 20 years.

The cause of breast cancer is not known but it has been of great attention to scientists for the last 30 years and therefore many risk factors have been established [6]. HRT was identified as a possible risk factor for breast cancer in the late eighties, but no clear and consistent association was demonstrated and published until this century. It was concluded that an increased risk of breast cancer was seen after 4-5 years of HRT exposure [7].

After the relationship between HRT and breast cancer had been shown the authorities withdraw the life-long osteoporosis indication. The relationship has however not been fully understood and their are many different opinions about the cause, different doses and specific hormones being more dangerous. The debate about HRT and breast cancer continues.

1.4 Prescription of HRT Today

Today, HRT is only indicated for treatment of moderate to severe menopausal symptoms and doctors are recommended to re-evaluate the need for continuing the treatment frequently. It is also recommended that treatment is individualised and that the women are treated with the lowest effective dose [8].

In the eighties the focus in the pharmaceutical industry was to develop continuous HRT preparations with lower doses than the existing sequential products. A large number of clinical trials investigated the relationship between estrogen dose and relief of menopausal symptoms as well as the safety profile.

It is well known that many women get satisfactory relief from the lowest available dose on the marked and the pharmaceutical industry has started development of ultra low doses. The introduction of these preparations will facilitate a more individualised HRT and in turn more women can be treated with lower doses.

2 The Bleeding Profile

2.1 Introduction

The "Bleeding Profile" is one of the most commonly used safety endpoints for this indication nevertheless no formal definition exist. For registration purposes the bleeding profile is a mandatory safety endpoint that should be studied in trials of at least 12 months according to both the European Medical Agency (EMEA) and the American authorities, the Food and Drug Administration (FDA) [9], [10].

No bleeding is referred to as amenorrhea. Furthermore EMEA guidelines recommend to study the combined endpoint of bleeding/spotting, spotting of course being inconvenient to the woman but rated as a less severe side-effect unless persistent.

The bleeding profile is usually recorded as a categorical variable with three levels: bleeding, spotting or no bleeding spotting. Bleeding can further be dived into: bleeding that require sanitary protection and bleeding that does not.

Bleeding is a self-assessment by the woman herself and it is usually recorded on a daily bases in some sort of diary that could be on paper or electronic. Electronic diaries have the advantage that they can be used to help reminding the woman to record data and that it can automatically transfer data to a central database. It is of great importance that the woman complete the diary regularly and that the information is reliable.

The disadvantage with this way of recording data is that the severity of the bleeding is not taken into consideration. Another disadvantage is that consecutive days of bleeding or spotting always are looked at as one episode, although at least for spotting it could very well be several episodes.

It is recommended that one distinguishes between genital bleedings and bleeding due to other causes such as bleeding due to a gynaecological examination or an endometrial biopsy. A known drawback of the continuous treatment is non-compliance with the medication, which can also cause bleeding or spotting. By comparing treatments in a randomised trial we hope to get a valid estimate of the treatment difference, however the description of the individual treatments may be trial specific and therefore misleading.

Even though the trial is randomised the aim should be to decide beforehand how the data should be described and analysed [11]. However, other relevant options for how to impute and exclude data should be investigated. A frequently used approach is for example to analyse the data including and excluding bleedings started in conjunction with a gynaecological examination to see how and if the results are affected. The point beeing that it is impor-

tant to fully understand the trial design and the data so that the analysis is fair and appropriate and valid descriptions of the data can be made.

In HRT trials it is necessary to characterize the bleeding profile differently depending on the treatment type. The reason for this is that the mechanism of the sequential treatment is to create a monthly bleeding whereas the mechanism of the continuous treatment is to avoid bleeding completely. For the continuous treatment, bleeding can therefore be thought of as a pure side-effect of the treatment, which should be thoroughly investigated if persistent. It is however not uncommon that a women experience bleeding during the first months of the continuous treatment.

2.2 Definitions of Derived Variables

2.2.1 The Proportion with at Least One Event

The proportion is defined as the percentage of women with at least one day of bleeding during a specific period. The proportion can be reported for different treatment periods such as: the complete trial, periods of three lunar months and for each separate lunar month. The proportion is often incorrectly referred to as an incidence, see for example [9].

2.2.2 Number of Events

An episode of bleeding is defined as a period of one or more consecutive days with bleeding, separated by at least one day of no bleeding. An episodes that starts in one period and continues on to the next should only be counted in the period that it starts. For the continuous treatment all episodes of bleeding can be regarded as irregular episodes, whereas for the sequential treatment one withdrawal bleeding per lunar month is expected.

2.2.3 Duration of Events

The number of days with bleeding in a period is used to give information on the duration of the event. It is however complicated to calculate the duration of bleeding episodes during a period. To do so one must decide if this endpoint should be calculated only for women with at least one episode and also consider if the duration of an episode that continues over two periods only should burden the period that it starts in. Since the interest is to assess how bleeding changes over time, the number of days of bleeding per time period appears to be both an intuitive and simpler endpoint to look at.

2.2.4 Time to First Event

The time to the first bleeding episode is calculated as days from treatment start. Women without any bleeding episodes are censored at their last day of treatment.

2.2.5 Proportion of Irregular Events

The proportion of irregular bleeding is only applicable for the sequential treatment since all episodes of bleeding for the continuous treatment could be referred to as irregular. The incidence is defined as the percentage of women with bleeding that occur before or after the withdrawal bleeding is expected. The withdrawal bleeding is expected to start in the third phase of the sequential treatment, i.e. after the phase which includes progestogen. Bleedings starting more than for example two days before or after this phase can be defined as irregular.

2.3 Bleeding or Bleeding/Spotting as the Event

The above endpoints can all be presented in terms of bleeding alone, bleeding/spotting or spotting alone. It is informative to describe both the incidence of bleeding alone and bleeding/spotting because they represent an adverse effect with increasing severity. This straightforward interpretation of increasing severity can however be misleading because a woman that experience both bleeding and spotting will have more days with bleeding/spotting but she may still have more bleeding episodes. An example is used to illustrated this: A women can have the following bleeding pattern (28 consecutive days), 0 means no bleeding, B means bleeding and S means spotting: 0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-0-B-B-S-B-S-S-0-0-0 This gives one episode of bleeding/spotting but two episodes of bleeding. Bleeding/spotting occur on 6 days whereas bleeding alone occur on only 3 days.

3 Statistical Analyses of the Bleeding Profile

The bleeding profile can be characterized in many ways and depending on the question there are many appropriate statistical models that can be fitted to the data. From a statistical viewpoint it is inadequate to specify the bleeding profile as a endpoint without clarifying how it should be assessed and derived [11]. Different treatments can be compared in terms of proportions, odds ratios, hazard ratios and rate ratios. Furthermore the treatment effect can be described using multiplicative or additive regression models depending on if we want to describe the hazard ratio or as an excess risk.

3.1 Data

To illustrate the statistical methods data for 200 women were selected. This data consist of daily diary recording for 100 women randomised to placebo and 100 randomised to active continuous HRT treatment. The event of interest is the combined endpoint of bleeding and spotting and the response can therefore be thought of as an adverse event. The distribution for the number of events per subject is illustrated in Figure 1. The aim of this essay is to illustrate the statistical methods and not the clinical use nor the effect per se of HRT.

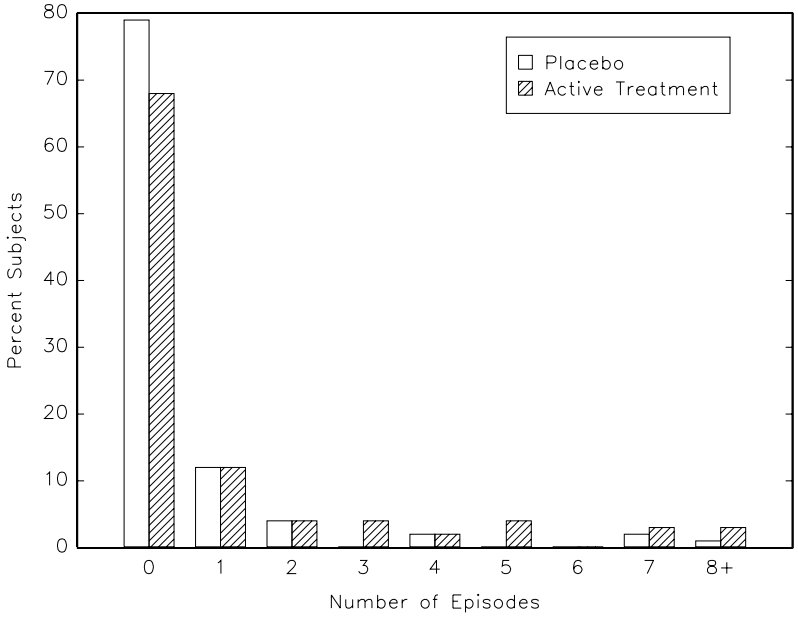

Figure 1: Distribution of Episodes per Treatment

3.2 Risk Factors for Bleeding

The risk factors associated with bleeding have been investigated in two other 6 month trials [12] and [13]. In [12] the conclusion was that women with a postmenopausal duration of 24 months or less, a pre-treatment endometrial thickness greater than 5 mm, and serum estradiol level greater than 25 pg/mL have increased risk to have endometrial bleeding within the first 6 months of continuous HRT. In [13] the risk of bleeding was increased for women with a postmenopausal duration of 24 months or less.

Publications on the subject of risk factors for the combined endpoint of bleeding and spotting are not easily available. The reason may be that this combined endpoint is more diluted than that of only bleeding which makes a potential association difficult to assess.

3.3 Baseline Characteristics

The relevant baseline characteristics for the trial data introduced in section 3.1 are given in Table 1. The table includes the known risk factors for bleeding, for which we have data in this trial, and the age at time of randomisation. The reason for including age is that the time since menopause may be difficult to measure, especially if many of the women are on birth control pills, that affect the menstruation, or if they have started HRT before entering the menopause.

Table 1: Baseline Characteristics

	Treatment Group									
	Placebo (n=100)					Active Control (n=100)				
Baseline Characteristics	Mean	SD	Median	Min	Max	Mean	SD	Median	Min	Max
Age (years)	56.2	4.9	56	47	65	54.9	4.5	55	45	64
Years since menopause	7.0	5.4	6	1	34	6.1	4.8	5	1	20
BMI (kg/m2)	25.0	3.4	24	18	35	25.3	3.5	25	21	35
Days in trial	173	10	171	153	222	174	8	174	157	206

The baseline characteristics were similar in the two groups, except for a slight shift in the age-distribution towards younger women in the active treatment group, see Figure 2. The different age distributions in the treatment groups become important when age is associated with the response. When age is ranked and divided into five groups of similar size the proportion of women experiencing at least one event decreases from 47, 40, 20, 18, 10% by increasing age-group. There is a clear trend towards lower proportions with increasing age. The linear relationship between age and proportion is illustrated in Figure 3, where the group proportion is plotted against the mean

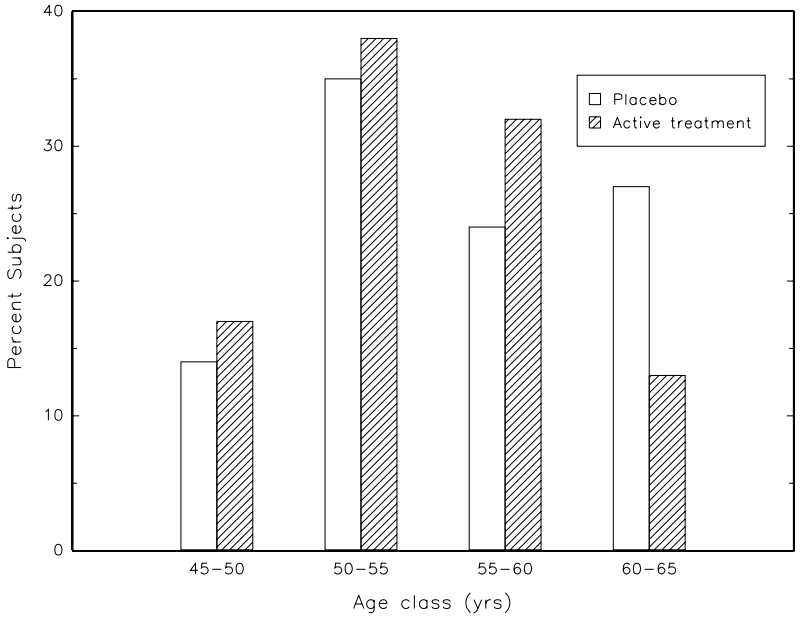

Figure 2: Age Distribution

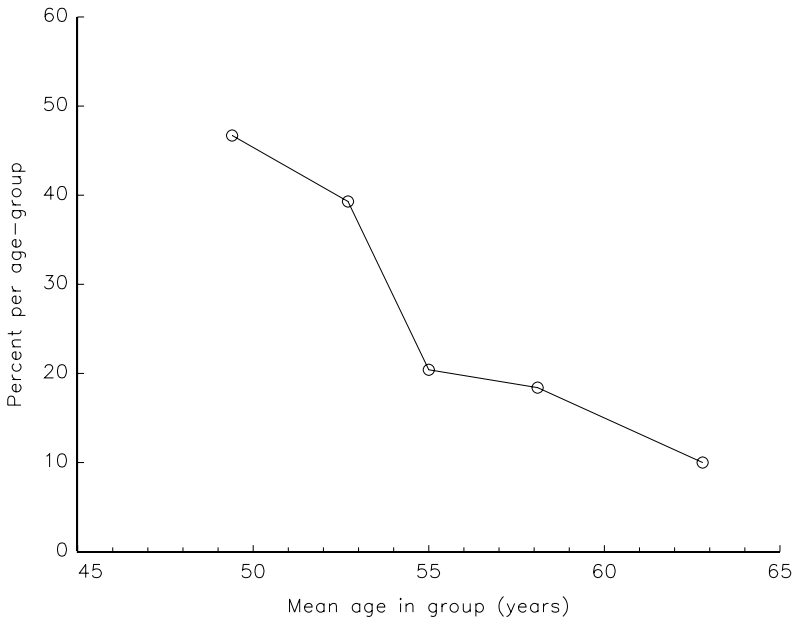

Figure 3: Proportion with Event per Age Group

age for the group. In the analyses also years since menopause and body mass index (BMI) were included. These variables do however not have an association with the combined endpoint of bleeding and spotting in this data material, so they will not be mentioned any further in this essay.

For the calculations presented hereafter age has been centred around the mean age for the group of included women. This is done to achieve meaningful interpretation of the intercepts for the models that allows for age adjustment. Treatment is coded as 1 for active treatment and 0 for placebo, so a zero vector of covariates gives a women of mean age in the placebo group.

3.4 Missing Data

In this trial we have complete data for all subjects. This is however a rare situation in most clinical trials and missing data can bias the treatment comparison. It is therefore necessary to compare the duration of treatment for the groups and to evaluate the reasons for discontinuation and non-compliance. It is recommended in the guideline [11] that the statistician ensures that the results are robust through sensitivity analyses. Here the potential outliers, missing data and protocol deviations should be taken into account. This topic will not be further discussed in this essay.

3.5 Proportions

The proportion of women having experienced at least one event during a period is a simple and straightforward endpoint. The proportion reduces the daily diary recordings to a simple binary outcome. However with this simple endpoint there is an issue about what to do with subjects that withdraw early from the trial without having experienced an event, since the observation time is not taken into account.

3.5.1 The Two by Two Table

The proportions should be described by presenting the number of subjects with an event divided by the total number of subjects for each group, p_0=P(event in placebo group) and p_1=P(event in active treatment group). The proportions can be compared using simple and well known statistical tests such as the Chi-Square Test or Fishers Exact Test.

3.5.2 Mantel-Haenszel and Logistic Regression

If we want to control for one or more discrete factors, the Mantel-Haenszel Chi-Square Test can be used and if we want to control for both discrete factors and continuous covariates a logistic regression model can be used. In both cases the odds ratio is used to describe the efficacy. The odds ratio (OR) is defined as

$$\text{OR} = \frac{p_1/(1-p_1)}{p_0/(1-p_0)}$$

and it is the odds for an event in the active treatment group relative to the reference group. The odds ratio is however not easily understood by most people and therefore it is a good idea to always also present the proportions.

In a logistic regression model we assume that, conditional on measured covariates, the probability of an event is the same for all subjects in a treatment group. More formally, we assume that subjects are independent and that n_0 and n_1 are the number of subjects with an event in the placebo and the active treatment group, respectively, with

$$n_i \sim \text{Bin}(N_i, p_i).$$

We model the dependence of p_i on treatment and age for a subject as a linear function of the logit for p_i. The following linear logistic regression model was fitted to the data

$$\text{logit}(p_i) = \log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 \cdot \text{treatment}_i + \beta_2 \cdot \text{age}$$

where treatment_i=1 for active treatment and 0 for placebo, β_0 is the intercept, β_1 and β_2 are the regression coefficients for treatment and age respectively [14].

The interpretation of the regression coefficients are as follows: β_0 is the logit for a women of mean age in the placebo group; β_1 is the log odds ratio for active treatment versus control and finally β_2 is the log odds ratio for an event if the age is increased with one year. If age is excluded from the model the interpretation of β_0 is changed to the logit for all women in the placebo group. Excluding age may however confound the effect of treatment on response and induce a spurious association.

3.5.3 Statistical Analyses of Proportions

The proportion of women experiencing at least one event during 6 lunar months of treatment are estimated to 21% in the placebo group and 32% in the active treatment group, see Table 2. When comparing the treatment

Table 2: Summary Table of Event using a Binary Variable

Group	No Events	At least one Event	Total
Placebo	79	21	100
Active Treatment	68	32	100
Total	147	53	200

Table 3: Statistical Analysis of Two by Two Table

Analyses	Model	Variable	Odds Ratio (Active/Placebo)	p-value
Chi-Square	Treatment			0.078
Fishers Exact Test	Treatment			0.109
Mantel-Haenszel	Only Treatment	Treatment	1.77	0.079
	Treatment and Age	Treatment	1.56	0.257
		Age vs. Event		<0.0001
Logistic regression	Only Treatment	Treatment	1.77	0.080
	Treatment and Age	Treatment	1.55	0.203
		Age	0.85	<0.0001

groups no statistically significant difference is found (p=0.078 with the Chi-Square Test and p=0.109 with Fishers Exact Test), see Table 3.

Without accounting for age the odds ratio for experiencing an event in the active treatment group compared to the placebo group is estimated to 1.77 (p=0.080 with the Likelihood Ratio Test for the logistic regression model and p=0.079 for the Mantel-Haenszel Chi-Square Test), see Table 3.

The different age distributions can be taken into account by fitting age either as a linear covariate or as a discrete factor in the logistic regression model or by stratifying for age in the Mantel-Haenszel Test. It appears reasonable to assume a linear relationship between age and the logit for an event as can be seen from Figure 4. Here, age has been divided into five groups of equal size and for each group the logit for an event is plotted towards the mean age for the group.

The event is highly associated with age when applying the models above. When the logistic regression model is fitted (see Table 3) the age adjusted odds ratio for an event in the active treatment group compared to placebo is 1.55 ($\exp(\beta_1)$ in the model). This is slightly smaller than for the un-adjusted odds ratio and still not significant (p=0.203). The regression coefficient for age, i.e. $\exp(\beta_2)$, is estimated to 0.85 (p<0.0001) and should be interpreted as the odds ratio for an event if age is increased by one year. Similar results are obtained with the Mantel-Haenszel where the odds ratio is estimated to 1.56 when age is divided into five groups. The estimated odds ratio is also similar when fitting age as a discrete factor in the logistic regression model (data not shown here). With fewer age-groups than five, with both methods,

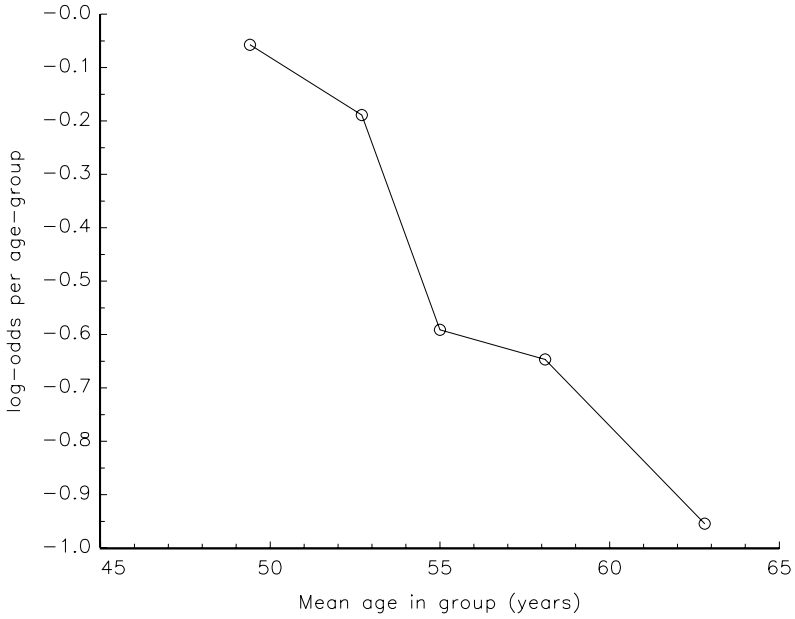

Figure 4: Log-Odds for Event per Age Group

the odds ratio is nearer to the estimate obtained when age is not controlled for, i.e. 1.77.

3.5.4 Conclusions - Proportions

During the 6 lunar months 21% in the placebo group and 32% in the active treatment group experienced at least one event. When bleeding is coded as a binary event it is highly associated with age and the trend appears to be decreasing linearly with age. Due to different age distributions in the two treatment arms it is necessary to control for age when comparing the treatments groups. When doing so, the age adjusted odds ratio is estimated to 1.55 with the logistic regression model and the age-stratified odds ratio to 1.56 with the Mantel-Haenszel, i.e. the age adjusted treatment effect is slightly lower than the crude treatment effect of 1.77. The odds ratio for the active treatment group compared to placebo is not statistically significantly increased (p=0.203) even though it is estimated to be 55% higher.

3.6 Time to First Event

By accounting for the time to the first event we can differentiate between groups that have the same proportion for an event but where one group experiences the event earlier than the other. Furthermore, a subject that withdraws from the trial without having experienced an event contributes to the risk population until the time of withdrawal, but no longer.

In section 3.6.1-3.6.3 we describe different statistical methods and models for time to event data, in section 3.6.4 we apply these methods to the trial data, with a summary in section 3.6.5.

3.6.1 Kaplan-Meier and the Log-Rank Test

The Kaplan-Meier estimator is a widely used, non-parametric, method to estimate the survival function, accounting for censoring (e.g. early withdrawals).

The most common way to compare survival functions is the non-parametric log-rank test. The Wilcoxon test is a weighted version of the log-rank test but more powerful to detect differences that occur at early time points [15]. The log-rank test allows us to control for discrete factors but not for continuous covariates. Furthermore it is only a test, and as such does not quantify the treatment effect, which is a limitation.

For the time to event or survival data, we are interested in estimating the treatment effect and do so by comparing the hazards. If we want to estimate the mean or the median time, where half of the subjects are event-free, we need to use a model where the distribution of the event-time is specified parametrically such as the exponential or the Weibull distribution. This is of course possible, but generally it is difficult to impose adequate parametric assumptions on the distribution of the event-time.

3.6.2 Cox Regression Model

In a Cox regression model we can account for both discrete factors and continuous covariates affecting the hazard, and we can quantify the treatment effect in terms of a hazard ratio. The instantaneous risk, or the hazard function, is denoted

$$\lambda(t) = \lim_{h \to 0} P(t \leq T < t + h | T \geq t)/h,$$

i.e. the density function for having an event at time t, given that no event occurred before time t. We write

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{dlogS(t)}{dt}$$

with $f(t)$ the (unconditional) density function and $S(t) = P(T > t)$ the survival function. The survival function, can be expressed in terms of the hazard function as

$$S(t) = e^{-\int_0^t \lambda(s)ds}$$

The following Cox regression model was fitted to the data

$$\lambda_i(t) = \lambda_0(t) \cdot e^{\beta_1 \cdot \text{treatment}_i + \beta_2 \cdot \text{age}}$$

where treatment$_i$=1 for active treatment and 0 for placebo, $\lambda_0(t)$ is a non-parametric baseline hazard function, β_1 and β_2 are the regression coefficients for treatment and age respectively.

The Cox regression model assumes proportional hazards over time, i.e. $\lambda_i(t) = \theta \cdot \lambda_j(\text{t})$. The proportional hazards assumption can be checked by using the relation

$$\log(-\log S_i(t)) = \log\theta + \log(-\log S_j(t))$$

indicating that the $\log(-\log S_i(t))$ should be parallel [15].

The Cox regression model can be checked for time varying effects through the newly developed resampling techniques for martingales [18]. A Cox model allowing for time varying effects differs from the above Cox model in that the regression coefficients are functions of time, i.e. β_i becomes $\beta_i(\text{t})$.

We assume that N(t) is a counting process that jumps when an event is observed. We can choose a compensator $\Lambda(t)$ so that N(t)-Λ(t) is a Martingale, i.e. a kind of a residual, with zero mean. It can be shown that this compensator is the cumulative hazards function, $\Lambda(\text{t})=\int_0^t \lambda(s)ds$. For the multivariate case, where we have independent and identically distributed counting processes, the variance of the cumulative hazard function can be estimated via the resampling techniques and used for statistical inference. This and other relevant functions can be used to assess proportional hazards, time varying effects and check goodness of fit.

3.6.3 Aalens Additive Model

The Aalen additive model is less commonly used and differs from the Cox model in that the treatment effect is estimated in terms of an excess risk.

The following Aalen additive model with time varying effects was fitted to the data

$$\lambda_i(t) = \beta_0(t) + \beta_1(t) \cdot \text{treatment}_i + \beta_2(t) \cdot \text{age}$$

where treatment_i=1 for active treatment and 0 for placebo, $\beta_0(t)$ is a non-parametric baseline hazard function, $\beta_1(t)$ and $\beta_2(t)$ are time varying effects for treatment and age respectively.

3.6.4 Statistical Analyses of Time to First Event

The time to the first event can be visualized by a Kaplan-Meier plot. It may also be illustrative to include the timing of the censored observations, see Figure 5.

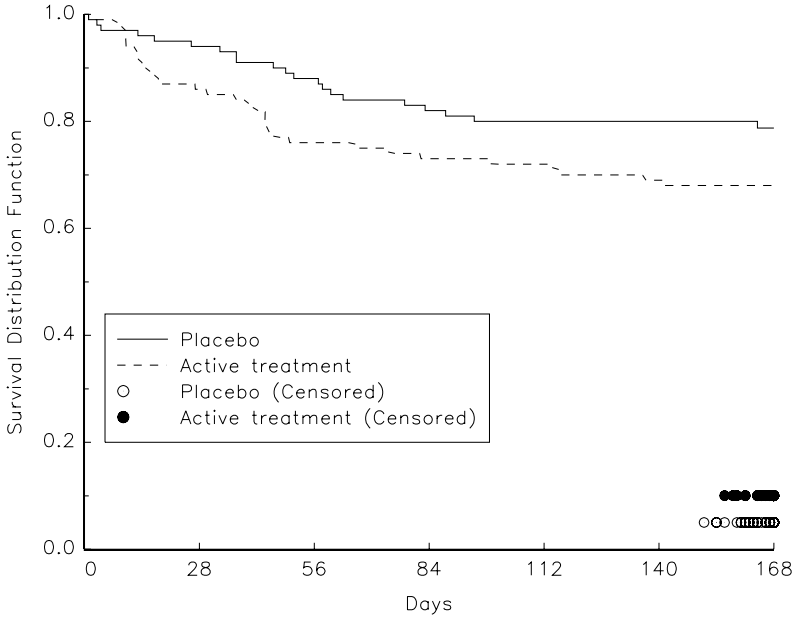

Figure 5: Time to First Bleeding/Spotting Episode

The statistical analysis of the time to the first event, not correcting for age, gives similar results as the previous analyses made on the proportion, see Table 4. The hazard ratio from the Cox model, i.e. $\exp(\beta_1)$, is estimated to 1.66. The age adjusted hazard ratio (1.51) is slightly lower but similar to the odds ratio in the logistic regression model. The hazard ratio for age, i.e. $\exp(\beta_2)$, is estimated to 0.88 ($p<0.0001$) and should be interpreted as the relative hazard for an event if age is increased with one year.

The Cox model was also investigated for time varying effects. The cumulative residuals are plotted against treatment and age in separate panels in Figure 6.

Table 4: Statistical Analysis of Time to First Event

Analyses	Model	Variable	Hazard Ratio	p-value
Log-Rank Test	Only Treatment	Treatment		0.069
	Treatment and Age	Treatment		0.055
Wilcoxon Test	Only Treatment	Treatment		0.058
Exponential	Only Treatment	Treatment		0.059
Cox Regression	Only Treatment	Treatment	1.66	0.073
	Treatment and Age	Treatment	1.51	0.145
		Age	0.88	<0.0001
	Treatment and Age(t)	Treatment	1.52	0.146
			Excess risk	
Aalen Model	Only Treatment	Treatment	0.0010	0.069
	Treatment and Age	Treatment	0.0008	0.149
		Age	-0.00025	0.0003
	Treatment and Age(t)	Treatment	0.0008	0.152

This plot is meant to give guidance to if the regression coefficients can be fitted as constants, i.e. β(t)= β. If the cumulative residuals are linear over time the coefficient can be fitted as a constant, this appears to be the case for treatment but not for age. However the Kolmogorov-Smirnoff and Cramer von Mises tests based on test processes both indicate that age can be fitted as a constant (p=0.216 and p=0.078) [18].

As can be seen in the Kaplan-Meier plot, the two survival functions cross one another after a few days, however after that the plot of the two $\log(-\log(S_i(t))$ functions are parallel, see Figure 7, indicating proportionality of the hazards.

To see if it is appropriate to fit age as a linear covariate in the Cox model, age has also been fitted as a discrete factor. These fits (different number of groups) gives similar hazards ratios for the treatment comparison as when fitting age as linear covariate (data not shown here).

The Aalen additive model was investigated by first fitting the model with time-varying effects, see Figure 8. As can be seen from this figure the cumulative residuals for age are not linear over time, and therefore it would be inappropriate to fit age with a constant factor (Kolmogorov-Smirnoff and Cramer von Mises tests gives p=0.028 and p=0.008). Treatment can however be fitted as constant. The excess risk is estimated to 0.000789 per day for active treatment compared to placebo, corresponding to an excess risk of 0.13 for 6 months of treatment.

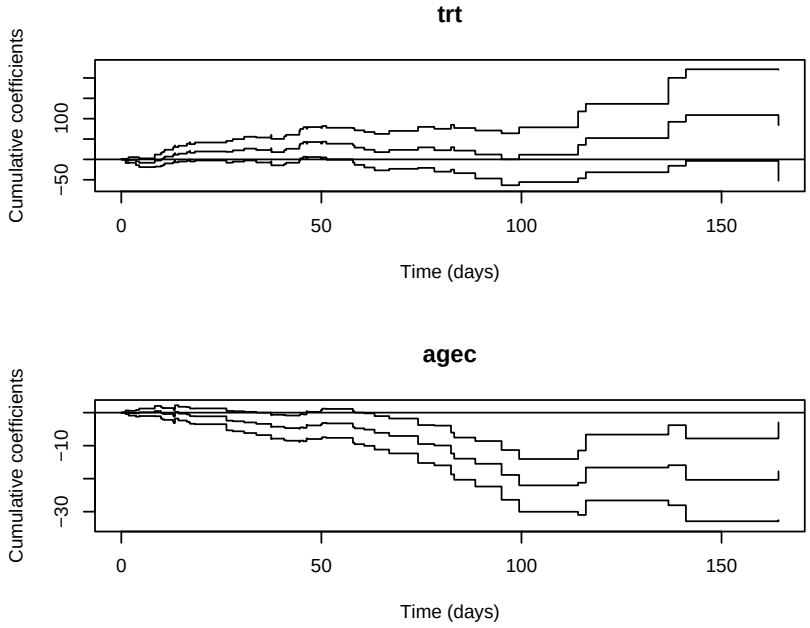

Figure 6: Observed Cumulative Residuals vs Continuous Covariates - Cox Model

3.6.5 Conclusions - Time to First Event

The Kaplan-Meier plot reveals two separated survival functions. When accounting for age in the Cox model the hazard ratio is estimated to 1.51 which is similar to that obtained for the logistic regression model. No statistically significant results is found when comparing the two groups (p=0.073). When fitting Aalen additive model age should be fitted as a time varying effect and the excess risk is estimated to 0.13 for 6 months.

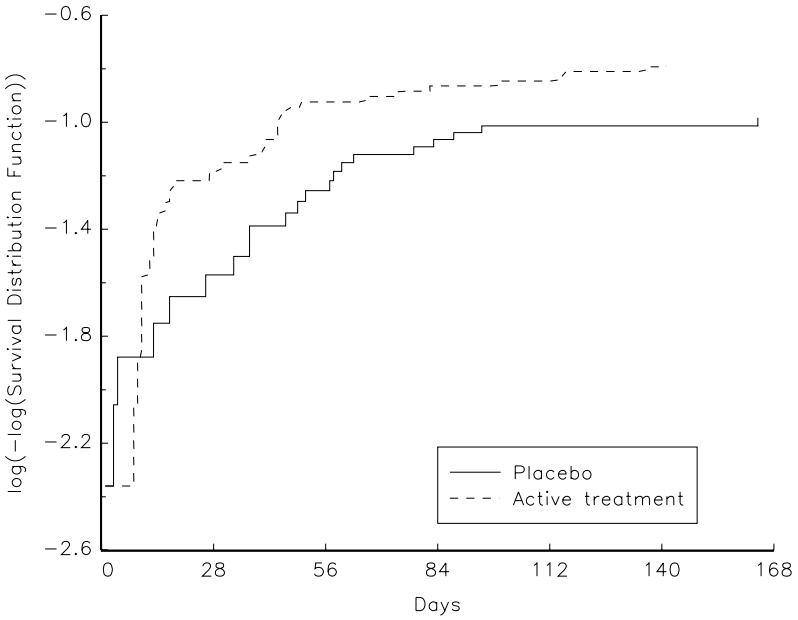

Figure 7: Log(-Log(Survival)) per Treatment Group

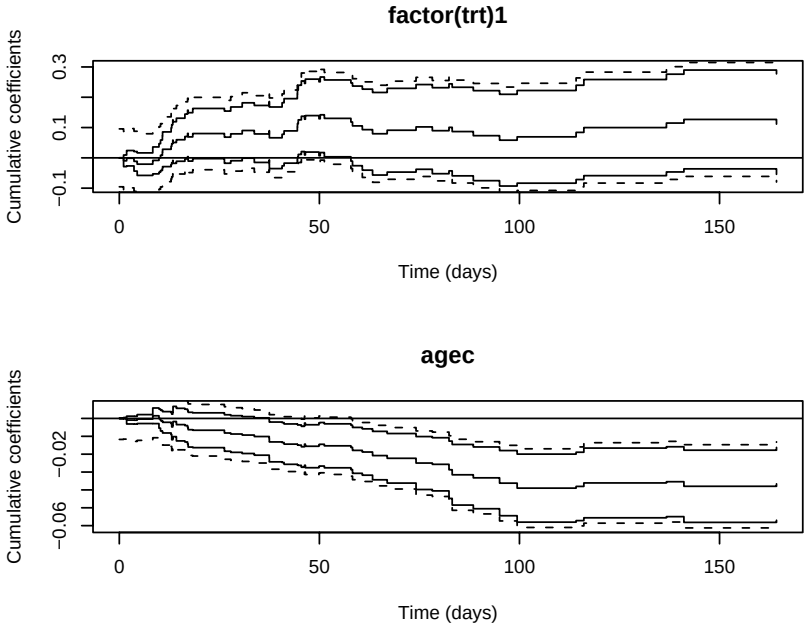

Figure 8: Observed Cumulative Residuals vs Continuous Covariates - Aalen Model

3.7 Rates

3.7.1 The Two by Three Table

As can be seen from the descriptive statistics there seems to be a tendency for more events per woman in the active treatment group than in the placebo group. The mean number of events per subjects is 1.1 (sd 2.2) in the active treatment and 0.5 (sd 1.4) in the placebo group. The median number of events is 0 in both groups and the range is 0-11 in the active and 0-8 in the placebo group. One could argue that the binary representation is too crude and that it would be relevant to include more than two groups. This can be achieved by splitting the 2 by 2 table into a 2 by 3 table, which will maintain acceptably large cell-sizes, see Table 5. We can then test if the cell proportions are equal for the two treatments groups using the following hypotheses:

$$H_0 : P_{\text{Active(k)}} = P_{\text{Placebo(k)}} \text{ for all } k = 1, 2, 3 \text{ against}$$
$$H_1 : \text{ not all are equal.}$$

This hypothesis can be tested using the Chi-Square test or the stratified Mantel-Haentzel test. These two tests treat the number of event categories differently, the Chi-Square has unordered categories while the Mantel-Haentzel has ordered categories [16].

Table 5: Summary Table of Event using an Ordered Categorical Variable

	Number of Events			
Group	0	1	2+	Total
Placebo	79	12	9	100
Active Treatment	68	12	20	100
Total	147	24	29	200

The Chi-Square test gives a non-significant result (p=0.082) on 2 degrees of freedom. In contrast a statistically significant association is obtained by the Mantel-Haentzel test (p=0.033) on 1 degree of freedom, see Table 6. Unfortunately, for tables larger than 2 by 2, the exact Mantel-Haentzel test has not yet been implemented in the software used for analysing this data (SAS version 9.1). The corresponding non-exact test is not appropriate for this data, since the cell counts for some age strata are very small (and even zero).

A more refined model that can be fitted to this type of data is the conditional logistic regression model [14].

Table 6: Statistical Analysis of the Two by Three Table

Test Statistic	DF	Value	p-value
Chi-Square	2	4.9955	0.082
Mantel-Haenszel Chi-Square	1	4.5270	0.033

3.7.2 Estimating the Event Rate

A simple way to estimate the event rate in a more sophisticated way than to calculate the mean number of events per subjects, is to calculate the number of events divided by the number of patient time units on treatment. In the active treatment group 109 events are seen during 17394 days of treatment giving a half-year rate of 1.1. Similarly in the placebo group 50 events are seen during 17259 days of treatment giving a half-year rate of 0.5. The reason for the rate being so similar to the mean number of events in our data is that almost all subjects completed the trial and therefore the exposure was similar in the two groups.

3.7.3 Poisson Regression and the Negative Binomial Model

Treatments can be compared using rates if we assume that, conditional on treatment group and measured covariates, the number of events per subject follows a Poisson distribution.

In a Poisson regression model we can assess how the rate depends on various discrete factors and continuous covariates, we can account for different observation times (usually fitted as an offset variable) and we can also allow for extra-Poisson variation (dispersion). The dispersion parameter is defined as a constant which is multiplied to the variance. Several different formulas are available for estimating this extra-Poisson dispersion constant [17].

If we assume that, conditional on treatment group and measured covariates, event rates for subjects in the Poisson model follow a gamma distribution, then the marginal distribution of the number of events follows a Negative Binomial distribution. The variability in the rate between subjects is accounted for by the extra parameter ϕ in the Negative Binomial model, $\mathrm{E(Y)}=\mu$ and $\mathrm{Var(Y)}=\mu+\phi\cdot\mu^2$. Similarly to the Poisson regression model, different observation times are handled as offset in the Negative Binomial model [17].

The following Poisson regression and Negative Binomial model was fitted to the data.

$$\log(\mu_i) = \log(t) + \beta_0 + \beta_1 \cdot \mathrm{treatment}_i + \beta_2 \cdot \mathrm{age}$$

where $\mathrm{treatment}_i$=1 for active treatment and 0 for placebo, t is the observation time (the offset), β_0 is baseline level, β_1 and β_2 are the regression

coefficients for treatment and age respectively. A logarithmic link function is used to obtain the rate ratio (or the relative rate ratio) for active treatment versus placebo.

Several models are fitted to the data. First both Poisson and Negative Binomial models are fitted without the logged observation time as the offset and without any covariates. In this case both methods give the same estimates for the rates and therefore also the same rate ratio. Inclusion of continuous covariates or the logged observation time generally gives higher estimates for the group rates with the Negative Binomial model than with the Poisson regression. Furthermore extra-Poisson variation is allowed, and the Poisson regression model is fitted both using the Pearson chi-square and the deviance to estimate dispersion [17].

3.7.4 Statistical Analysis of Rates

When applying Poisson and Negative Binomial regression to the data, excluding age, all p-values for the treatment comparison are below 5%, see Table 7. However, when including age as a continuous covariate, treatment is no longer significant in many of the analyses. Statistically significant differences are obtained with the Poisson model, except when estimating over-dispersion with the Pearson formula (p=0.078). The p-value for the treatment comparison from the Negative Binomial model remained not significant (p=0.113).

Table 7: Statistical Analysis of Event Rates

Analyses	Model	Offset	Rate Ratio Active/Placebo	Over-Dispersion	p-value
Poisson	Only Treatment	No	2.18	No	<0.0001
				Pearson	0.027
				Deviance	0.004
		Yes	2.16	No	<0.0001
				Pearson	0.027
				Deviance	0.005
	Treatment and Age	No	1.98	No	<0.0001
				Pearson	0.078
				Deviance	0.011
		Yes	1.96	No	<0.0001
				Pearson	0.078
				Deviance	0.012
Negative Binomial	Only Treatment	No	2.18	-	0.030
		Yes	2.17	-	0.030
	Treatment and Age	No	1.82	-	0.109
		Yes	1.80	-	0.113

The rate ratio for an event in the active group compared to placebo is

achieved by exponentiation of β_1. The rate is almost doubled in the active group compared to placebo, with an estimate of 1.96 from the Poisson and 1.80 from the Negative Binomial model. The rate ratio is generally a little higher when not accounting for the offset. For the Poisson model quite different results are obtained for the treatment comparison depending on the method used to estimate over-dispersion. The Pearson formula impose the most variance and therefore a non-significant result. The models where the dispersion parameter was not adjusted for or where the Deviance formula was used, gave very small p-values for the treatment comparison.

The intercept in the model can be interpreted as the rate ratio for a women of mean age in the placebo group ($\exp(\beta_0)$) and it is estimated to exp(-0.69)=0.5 in the models without offset. This estimate corresponds very well to the previously presented mean number of events and the half year rate, both also 0.5 in the placebo group.

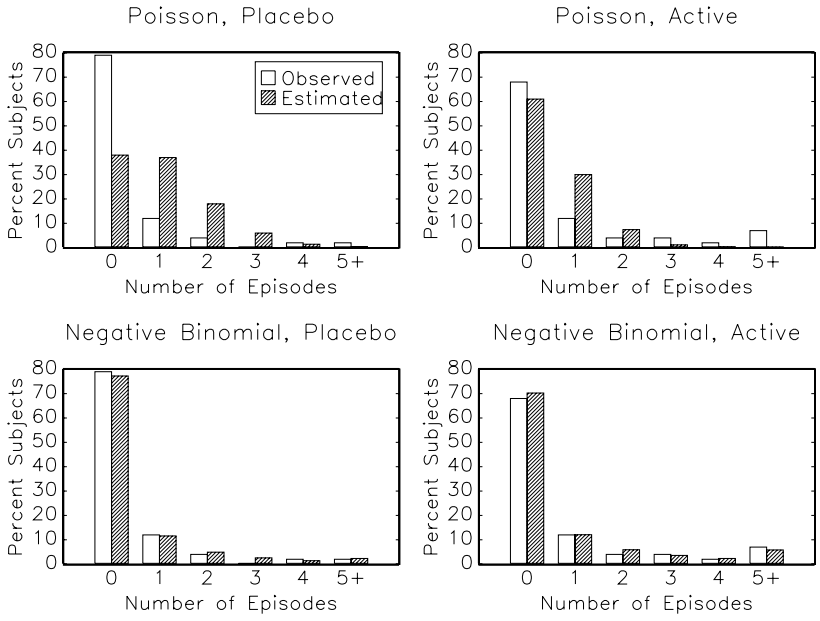

Figure 9: Observed vs. Estimated Distributions for Poisson and Negative Binomial Models

A common problem with the Poisson model is that it usually does not fit the data very well. Typically we have both too many subjects with no event and a few subjects with very many events and this cannot be accounted for in this model. The ratio of the deviance and the degrees of freedom (Df) is a measure of fit for theses models and it should generally not be above 2. For the Poisson model it is never below 2.4, clearly indicating lack of fit. The fit of the Negative Binomial is much better with deviance/Df=0.6. To illustrate the lack of fit, the actual number of events are plotted versus the expected

number of events for both the Poisson and the negative binomial model, see Figure 9.

3.7.5 Conclusions - Rates

The event rate is more than doubled in the active group (1.1) compared to the placebo group (0.5). The fit of the Negative Binomial model is superior to that of the Poisson regression (Figure 11) for these data. Adjusting for age, the rate ratio (Active/Placebo) is estimated to 1.80 and the corresponding p-value for the treatment comparison is not significant (p=0.113).

3.8 Repeated Time to Event

The time to the first, second, third etc. event can be compared by using extended Cox regression models. The event rate can be compared between groups by estimating the relative risk of having an event in the active treatment group compared to placebo. There are also several different plots that can be made to visualise the frequency and onset of events.

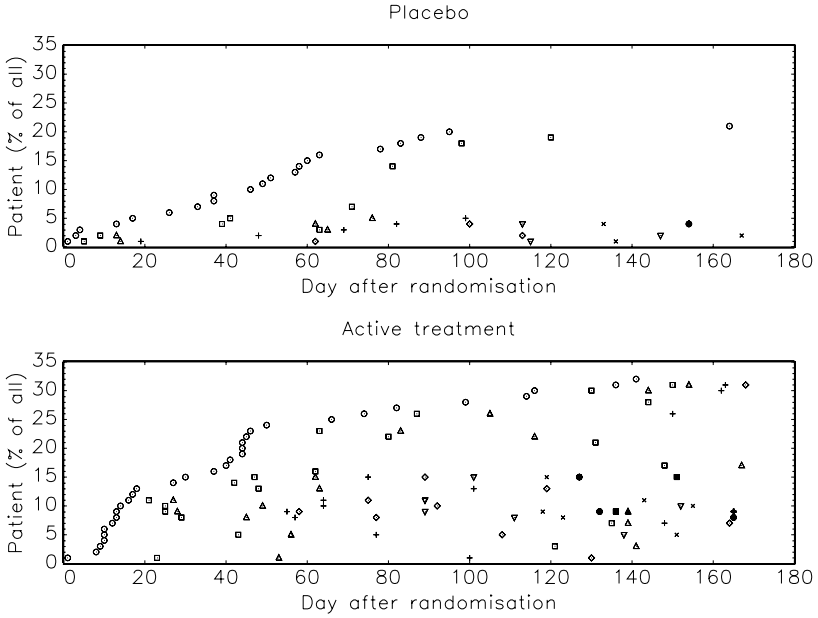

Figure 10: Time to Recurrent Events

Figure 10 shows how event occurrence over time can be presented. Each value on the y-axis represents one patient, the x-axis is the time scale and each

symbol represents one event. Patients are sorted according to first event time. The event number is displayed using different symbols. The curve formed by the maximal y-values thereby represents the cumulative number (or percentage) of patients having an event at that time-point without adjusting for withdrawals.

The next plot is an extension of the cumulative hazard, see Figure 11. This plot illustrates the mean number of events per patients over time and can be calculated by first calculating the expected number of events at each actual event-time and then sum this cumulatively. The expected number of events at an event time is the number of events at that time divided by the number of patients at risk. Patients at risk are patients not yet censored and not currently having an event. This plot is simple and easy to understand and gives a good feeling for how the treatment effect varies over time.

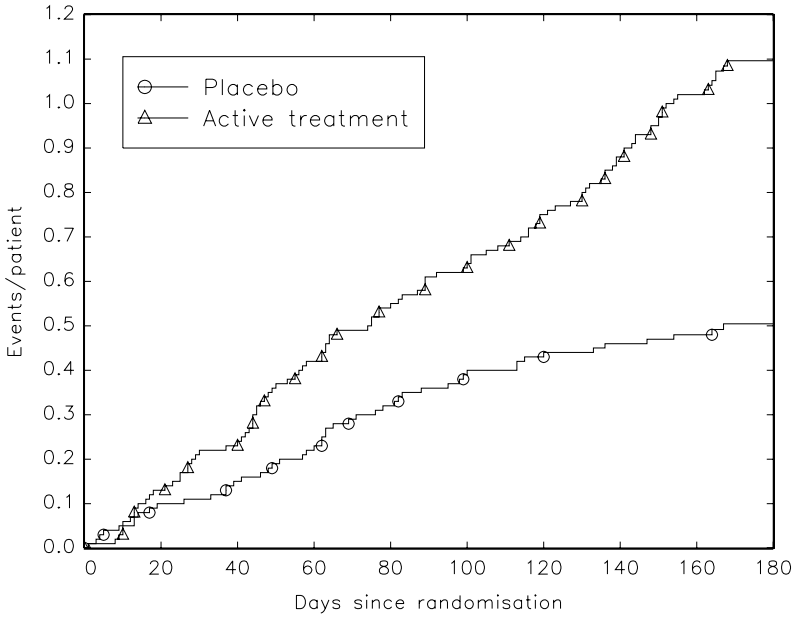

Figure 11: Mean Number of Events per Patient vs. Time

The last plot is a combination of several reversed Kaplan-Meier plots (1-Survial function), see Figure 12, displaying time to the first event, to the second event and so on. The aim of this plot is again to visualise the frequency and the timing of events. In this case, the last few panels do not include so many events, but they may still be regarded as illustrative.

The episodes can be analysed as recurrent events using a gamma frailty model. This model is an extended Cox regression model with treatment and baseline age as covariates, extended with a random effect (following a gamma distribution), which acts multiplicatively on the baseline hazard

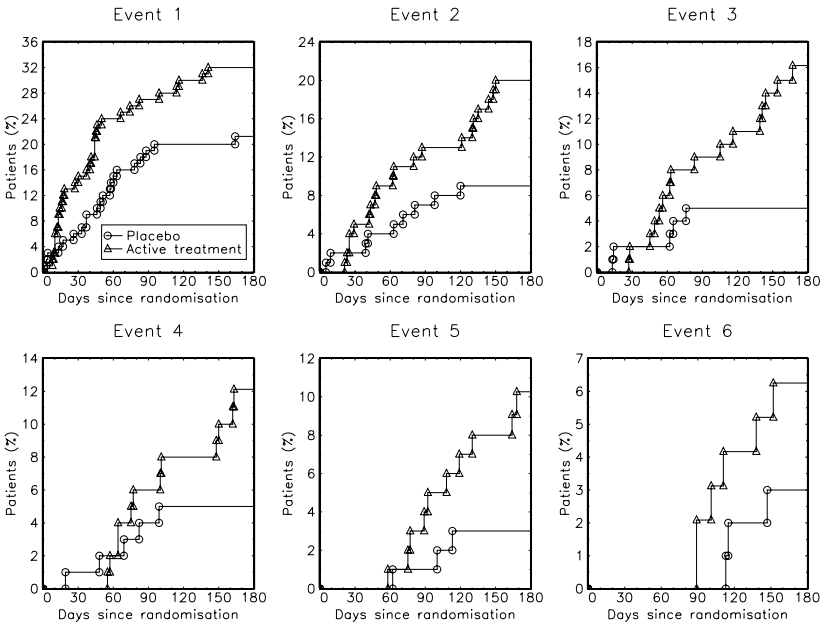

Figure 12: Time to Event No. 1-6

function describing the individual risk (or frailty) for a subject.

The following Gamma Frailty model was fitted to the data

$$\lambda_i(t) = z_j \cdot \lambda_0(t) \cdot e^{\beta_1 \cdot \text{treatment}_i + \beta_2 \cdot \text{age}}$$

where treatment_i=1 for active treatment and 0 for placebo, z_j the gamma(δ,θ) distributed frailty, representing random effect for each subject with δ=θ and therfore mean=1, $\lambda_0(t)$ is a nonparametric baseline hazard function, β_1 and β_2 are the regression coefficients for treatment and age respectively [19].

3.8.1 Statistical Analyses of Repeated Time to Event

The results from fitting the frailty model is in line with the results from fitting the Negative Binomial model. If we fit a model without age, the hazard ratio is estimated to 2.18 and it is statistically significantly increased for the active treatment versus placebo (p=0.028). If we include age the hazard ratio is reduced to 1.82 and the treatment comparison is no longer significant (p=0.110).

Table 8: Statistical Analysis of Repeated Time to Event

Analyses	Model	Variable	Hazard Ratio	p-value
Frailty Model	Only Treatment	Treatment	2.18	0.028
	Treatment and Age	Treatment	1.82	0.110
		Age	0.94	0.112

3.8.2 Conclusions - Repeated Time to Event

Several different plots can be used to illustrate the frequency and timing of events. These plots are easy to understand and informative. The hazard ratio for a women of mean age is estimated to 1.82. The similar hazard ratio estimated with the Cox regression model is 1.51, which indicates that more information is provided by including all subsequent events and not only the first.

4 Regulatory Guidance

Both the European Medical Agency (EMEA) and U.S. via the Food and Drug Administration (FDA) and are developing new guidelines for investigational studies for HRT [9] and [10]. Both these guidelines are currently only drafted and they recommend that new preparations with lower doses of combined estrogen and progestogen should focus on providing endometrial safety while efficacy is maintained and the bleeding profile is acceptable.

FDA proposes placebo-controlled studies and that the lowest effective dose is identified. The problem to obtain endometrial safety with the lowest effective progestogen dose is itself a challenge but it will not be further discussed in this essay.

The drafted guidelines are supported by the pharmaceutical industry, however there are some statistical issues that needs to be discussed with the authorities. Furthermore there are many ways for how the statistical hypotheses can be defined and how a test strategy can be set up. This section is written to give guidance for how an effective regulatory strategy can be translated into the statistical framework.

4.1 Efficacy Variables

Efficacy is measured in terms of hot flushes and the draft FDA definition is as follows [10] "Vasomotor symptoms in postmenopausal women are commonly known as hot flushes or hot flashes. The severity of vasomotor symptoms are defined clinically as follows:

- Mild: sensation of heat without sweating
- Moderate: sensation of heat with sweating, able to continue activity
- Severe: sensation of heat with sweating, causing cessation of activity.

Subjective measures (e.g., daily patient diary entries) can be used as primary efficacy endpoints".

The following endpoints are usually derived from the daily recordings of hot flushes:

- number of moderate to severe hot flushes per week
- number of severe hot flushes per week

- the total number of hot flushes per week.

Examples of endpoints that aims to join the frequency and severity of all episodes during a week into a combined endpoint is the Hot Flushes Weekly Weighted Score (HFWWS), defined as:

$$HFWWS = 1 \cdot \text{No. mild} + 2 \cdot \text{No. moderate} + 3 \cdot \text{No. severe}$$

An easy modification that only address the moderate and severe episodes during a week is the

$$modHFWWS = 2 \cdot \text{No. moderate} + 3 \cdot \text{No. severe}$$

Responders can be defined from these endpoints such as "women with at least 90% improvement in the HFWWS".

The FDA recommends the Severity Score (SS), also in two versions, defined as:

$$SS_1 = \frac{2 \cdot \text{No. moderate} + 3 \cdot \text{No. severe}}{\text{No. of moderate and severe episodes}}$$

$$SS_2 = \frac{1 \cdot \text{No. mild} + 2 \cdot \text{No. moderate} + 3 \cdot \text{No. severe}}{\text{Total No. of episodes}}$$

These endpoints are less intuitive, but as long as the total number of hot flushes are reduced, they also work as endpoints assessing change in severity.

Several questionnaires have been developed to assess the wide range of symptoms characterizing menopause, some also try to assess what is commonly known as "quality of life". Example of such questionnaires are the Greene Climacteric Score and the Kupperman Menopausal Index.

4.2 The Draft FDA Guidelines Concerning Efficacy

The following endpoints are defined "For the treatment of moderate to severe vasomotor symptoms, we recommend the following co-primary endpoints:

- Mean change in frequency of moderate to severe vasomotor symptoms from baseline to week 4
- Mean change in frequency of moderate to severe vasomotor symptoms from baseline to week 12
- Mean change in severity of moderate to severe vasomotor symptoms from baseline to week 4

- Mean change in severity of moderate to severe vasomotor symptoms from baseline to week 12.

For estrogen alone products intended to treat moderate to severe vasomotor symptoms, we recommend that the primary efficacy analyses show a clinically and a statistically significant reduction, within 4 weeks of initiation of treatment and maintained throughout 12 weeks of treatment, in both the frequency and severity of hot flushes in the treated groups compared with the control groups."

4.3 Comments to Draft FDA Guidelines

The FDA recommends to study two efficacy variables, each with a repeated evaluation at two different time-points. A new drug should be effective in treating hot flushes so the suggested endpoints covering both frequency and severity appears logical at first glance. The choice of the time-points are also reasonable since it may take a while for a systemic drug to fully reach its clinical effect (not longer than 4 weeks) and that the effect should be persistent.

For ethical reasons a pharmaceutical company should try to minimise the number of subjects exposed to experimental drug [22]. This can be done by selecting a smart design and to choose endpoints that can be analysed with powerful statistical models. In turn, this requires previous knowledge of the endpoints and that different models have been fitted and compared. For this indication the parallel group design is more appropriate than a cross-over design because the duration of treatment should be 12 months and a long washout would be required. Furthermore, it is recommended to only include women with frequent and severe symptoms, and these symptoms are expected to decrease with time. The trial should be randomised and double-blind and include a placebo group as reference. Several doses should be included so that the lowest effective dose can be found.

The number of subjects to be included in the trial is usually based on "a detectable difference" which can be motivated, a variance estimate, a level of significance and the wanted power for a the trial. The level of significance is usually set to 5%, the power to between 80-95% and a good guess of the variance can usually be obtained from the literature or from previous trials.

4.3.1 Multiple Primary Endpoints

If we have one hypothesis, the level of significance is the probability that we falsely reject a true hypothesis. If we have several primary hypotheses we can define the overall alpha (sometimes called the global alpha or the

familywise alpha) as the probability that we falsely reject at least one of the hypotheses. The overall alpha is however not applicable in this case, we have four hypotheses and we need to win on all four to get a conclusive result for the trial.

It is natural that we assume that our four endpoints are correlated and that the endpoints looking at two different time-points should be highly correlated if the drug is effective. Since it is difficult to estimate the correlations between the four endpoints in advance our approach will most likely be to over-power the trial based on the least effective endpoint. The more correlated our endpoints are the less the power is lost for the "win on all four test scenario".

Other ways to interpret the guidelines are to ignore the fact that four endpoints recommended as co-primary endpoints and just look at the two endpoints that addresses efficacy after 4 weeks (which still raises the same problem but with 2 tests instead of four). One could also choose to consider one of the endpoint as the key primary and the other as secondary [11], however this is clearly not the intention in this case. A third alternative is to define a combined endpoint that takes care of both the frequency and the severity and get the authorities to accept this approach.

4.3.2 Multiple Comparisons

If we have more than one dose of the investigational drug, there is a multiplicity issue, since we have more than one direct treatment comparison per endpoint. If we have several doses that we want to compare to placebo in order to determine the lowest effective dose, we can benefit from pre-specifying a test strategy. Since the dose response relationship is well known, increasing efficacy and side-effects with increasing dose, the following hierarchical test strategy should be applicable for each efficacy endpoint. Start by comparing the highest dose to placebo and if a significant result is obtained, continue with the next highest dose etc.. Only continue testing if a significant result is obtained on the previous dose level. If this strategy is predefined it will solve the multiplicity issue.

Other ways of solving this multiplicity issue, which does not take the dose response relationship into account are often seen, but should be avoided since they are less powerful for this situation. Examples of such solutions are:

- use a overall test first. If this test is significant the pairwise tests can be presented.
- correct the p-value for multiple testing by using a special method such as Dunnet.

4.4 Draft EMEA Guidelines on Bleeding Control

For combined treatments, bleeding data should include in general bleeding or spotting, incidence of amenorrheic cycles (total absence of bleeding) and percentage of women with withdrawal bleeding where appropriate. More specifically:

For cyclic or sequential products:

- Bleeding data should include percentage of women with regular withdrawal bleeding, the mean duration of these bleedings, and the time they start before/after the last pill of the progestogen phase. Data should also include percentage of women with breakthrough bleeding and/or spotting appearing during the first three months and during months 10 to 12 of treatment. The incidence of amenorrhoea (no bleeding or spotting) during the first year of treatment should also be specified.

For continuous combined products:

- Bleeding data should include the incidence of amenorrhoea (no bleeding or spotting) during months 10 to 12 of treatment, and the percentage of women with bleeding and/or spotting appearing during the first three months of treatment and during months 10 to 12 of treatment.

4.5 Comments to Draft EMEA Guidelines

EMEA recommends several endpoints that should be evaluated at several time-points. However, these are safety endpoints and as such a description of the data is usually sufficient. No formal analysis is requested and if analyses are performed, there should be no correction for multiple endpoints. The requested endpoints make sense and can assess the bleeding profile over time. If a drug is compared to a competitor it is beneficially to use the most powerful statistical model, similar to what I have recommended in the previous chapter.

5 Summary and Discussion

HRT is effective for treatment of moderate to severe menopausal symptoms. Due to the association with breast cancer, women should be treated with the lowest effective dose and the need for treatment should be frequently re-evaluated by the doctor.

In this essay, the safety of the HRT treatment is evaluated by monitoring the endometrial safety through the bleeding profile. From the bleeding profile several endpoints are defined and analysed using the data from a clinical trial where active treatment is compared to placebo.

In the data used to illustrate a number of statistical methods and models, there is a small shift in the age distribution. Age is an important factor with regard to bleeding but not for the combined endpoint of bleeding/spotting [12] and [13]. However, in the data used here, age was consistently highly associated to the frequency and timing of the event of interest.

The analysis of repeated events is more powerful than the analyses that use a binary outcome or time to first event. For the more powerful methods it was necessary to include age in the model in order to get sharp results. It is of great importance to investigate the data with easy and straightforward methods and to get a good understanding of the data before applying the more complex methods. Several suggestions have been given for how to present and illustrate this type of data.

The large number of methods available for studying goodness of fit have been left out of this essay. It is of course necessary to check the model thoroughly and ways to do so are described in most text books. The topic is outside the scope for this essay.

A draw-back of the models presented here is that the duration of the event has not been taken into consideration. To do so, one could for example fit a repeated measurement model to the daily recordings. Such models include modelling of the covariance structure [20] and [21].

A number of experts on this subject have written guidelines (currently only drafted) which are intended for the pharmaceutical industry. How the bleeding profile should be looked at is part if these guidelines and therefore also discussed and addressed in this essay. Furthermore, a test strategy is suggested for the still rather unusual situation where we have several primary endpoints and where we need to show effect on all in order to achieve a positive result for the trial. A good understanding of the data, use of effective designs and powerful analyses are important for clinical trial evaluation.

References

[1] Farrell E. (2003). Medical choices available for management of menopause. *Best Practice & Research Clinical Endocrinology & Metabolism*, 17 (1), 1-16.

[2] Rödström K., et all. (2002). A longitudinal study of the treatment of hot flushes: the population study of women in Gothenburg during a quarter of a century. *Menopause*, 9 (3), 156-161.

[3] Barrett-Connor E. (2002). Hormones and the health of women: past, present and future. *Menopause*, 9 (1), 23-31.

[4] Parkin D. M. (2001). Global cancer statistics in the year 2000. *Lancet Oncology*, 2, 533-543.

[5] Socialstyrelsen. (2004). *Swedish Breast Cancer Statistics*. Available directly at http://192.137.163.40/EPCFS/. Accessed 25 September 2006.

[6] Key T. J., Verkasalo P. K., Banks E. (2001). Epidemiology of breast cancer. *Lancet Oncology*, 2, 133-139.

[7] Collaborative Group on Hormonal Factors in Breast Cancer. (1997). Breast cancer and hormone replacement therapy: collaborative reanalysis of data from 51 epidemiological studies of 52 705 women with breast cancer and 108 411 women without breast cancer. *Lancet*, 350, 1047-1059.

[8] Position Statement (2004). Recommendations for estrogen and progestogen use in peri- and postmenopausal women: October 2004 position statement of The North American Menopause Society. *Menopause*, 11, 589-600.

[9] Committee for medical products for human use. (2005). Guideline on clinical investigation of medicinal products for the treatment of hormone replacement therapy. EMEA, Draft. 20 January 2005.

[10] Guidance for Industry. (2003). Estrogen and Estrogen/Progestin Drug Products to Treat Vasomotor Symptoms and Vulvar and Vaginal Atrophy Symptoms - Recommendations for Clinical Evaluation. Draft Guidance. U.S. Food and Drug Administration, January 2003.

[11] ICH Harmonised Tripartite Guideline. (1999). Statistical Principles for Clinical Trials. *Statistics in Medicine*, 18, 1905-1942.

[12] Shau WY. (2002). Factors associated with endometrial bleeding in continuous hormone replacement therapy. *Menopause*, 9 (3), 188-194.

[13] Johnson J.V., et all. (2002). Postmenopausal uterine bleeding profiles with two forms of continuous combined hormone replacement therapy. *Menopause*, 9 (1), 16-22.

[14] Collett D. (2003). *Modelling Binary Data*. 2nd ed. Chapman & Hall, London.

[15] Collett D. (2003). *Modelling Survival Data in Medical Research*. 2nd ed. Chapman & Hall, London.

[16] Mantel N. (1963). Chi-square tests with one degree of freedom; extensions of the Mantel-Haenzel procedure. *Journal of the American Statistical Association*, 58, 690-700.

[17] McCullagh P., Nelder J.A. (1989). *Generalized Linear Models*. 2nd ed. Chapman & Hall, London.

[18] Martinussen T, Scheike T.H. (2006). *Dynamic Regression Models for Survival Data*. Springer Science, New York.

[19] Hougaard P. (2000) *Analysis of Multivariate Survival Data*. Springer Verlag, New York.

[20] Diggle P.J., Heagerty P., Liang KY, Zeger S.L., (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford.

[21] Molenberghs G, Verbeke G. (2005). *Models for Discrete Longitudinal Data*. Springer Science, New York.

[22] Declaration of Helsinki. (1997). Recommendations guiding medical physicians in biomedical research involving human subjects. *JAMA*, 277 (11), 925-926.