Does estrogen predict cognitive function in women at risk for breast cancer?

Feng Zhu

Department of Statistics

University of British Columbia

November 4, 2010
1 Introduction

This project is a secondary analysis of a double-blind randomized controlled trial of women who underwent chemoprevention treatment with a drug called Azoxifen. The project aims to examine thirteen predictors of cognitive function in women at risk for breast cancer. Some of the predictors and all the response variables were measured over three time points, and some of the predictors were recorded only at the beginning of the trial. The most important predictor is estrogen, and the treatment effect of drug will be treated as a moderator.

2 Data Description and Collection

2.1 Variables

A total 115 patients were divided into two groups, an experimental group and a control group. For all the patients, 10 indices of cognitive function, as dependent variables, were measured at three time points (baseline, 6-month, and 12-month). Five predictors were also measured at these three time points. Another eight predictors, such as age, years of education, etc, were recorded only at the baseline. There are three kinds of variables: binary, continuous, and sum of ordinal scales (which I think that you are comfortable to treat as continuous variables). The details for these variables are listed in Table 1.

All patients had the baseline measurements, but there were some patients who were lost to follow-up after the baseline measurements.
<table>
<thead>
<tr>
<th>Variables</th>
<th>Values</th>
<th>Type</th>
<th>Change over time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predictors</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>Continuous</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Education (years)</td>
<td>Continuous</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>Married or common law</td>
<td>Yes/No</td>
<td>Binary</td>
<td>No</td>
</tr>
<tr>
<td>Employment</td>
<td>Yes/No</td>
<td>Binary</td>
<td>No</td>
</tr>
<tr>
<td>Professional status</td>
<td>Yes/No</td>
<td>Binary</td>
<td>No</td>
</tr>
<tr>
<td>Caucasian ethnicity</td>
<td>Yes/No</td>
<td>Binary</td>
<td>No</td>
</tr>
<tr>
<td>Menopausal status</td>
<td>Pre/Post</td>
<td>Binary</td>
<td>No</td>
</tr>
<tr>
<td>Body mass index</td>
<td>Continuous</td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>CES-D depression</td>
<td>Sum of Ordinal</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>SAI state anxiety</td>
<td>Sum of Ordinal</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>SF36 physical health</td>
<td>Sum of Ordinal</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Estrone</td>
<td>Continuous</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Bioavailable E2</td>
<td>Continuous</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Responses</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Variables</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WMS auditory immediate</td>
<td>Sum of Ordinal</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>WMS visual immediate</td>
<td>Sum of Ordinal</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>WMS immediate memory</td>
<td>Sum of Ordinal</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>WMS auditory delayed</td>
<td>Sum of Ordinal</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>WMS visual delayed</td>
<td>Sum of Ordinal</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>WMS auditory recognition delayed</td>
<td>Sum of Ordinal</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>WMS general memory</td>
<td>Sum of Ordinal</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>MS working memory</td>
<td>Sum of Ordinal</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>FAS verbal fluency</td>
<td>Sum of Ordinal</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Cognitive difficulties</td>
<td>Sum of Ordinal</td>
<td>Yes</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Basic Description of Variables
2.2 Treatment

No patient received the drug before the baseline measurement. The patients in the experimental group took the drug right after the baseline measurement until 12 months. The patients in the control group took the drug right after the 6-month measurement until 12 months.

Since all the patients took the drug after the 6-month measurement, it is possible that the trial was not blinded anymore after 6 months. The possible reason is that the drug may already show some positive effect at the 6-month measurement, and the control group also received the treatment after the 6-month measurement.

2.3 Preliminary Work

You mentioned that you have done some preliminary analyses, examining the predictors of cognitive function both at the baseline (before application of the drug) and longitudinally before our initial meeting. Along with the brief description of the project, a table of descriptive statistics for all the variables was provided.

3 Statistical Issues or Questions

After the initial meeting, we understood the main question is whether estrogen predicts the cognitive function in women at risk for breast cancer. There are other statistical issues that were mentioned at the meeting. We summarize the questions as the following:

1. You proposed use of the linear mixed effects model and would like to carry out the analyses in SPSS. However, you are not sure how to deal with the five predictors
that were measured repeatedly over time. Also, you would like to know the most appropriate statistical method for analyzing these longitudinal data.

2. You would like to verify whether the randomization of the trial was effective.

3. You would like to incorporate the treatment with Arzonxifen (Drug) as a moderator in the analyses.

4 Proposed Statistical Methods

4.1 Preliminary Examination of Data

Since you prefer to analyze the outcome variables individually and are willing to treat these outcome variables as continuous variables, we suggest first plotting the outcome variables against each other at baseline and calculating their correlations to check the relationships between their variables (in the absence of any impacts of the treatment). If two of the outcome variables are highly correlated, there is little point to fitting models for each of them as the two outcome variables are essentially the same characteristic of the patients and results from the models will be very close. In such a case, you might only want to pick one variable from each group of highly correlated outcome variables to fit the model. If all the variables are highly correlated, you could simply pick one of the outcome variables or you might want to use the average of these ten outcome variables, for the detailed model fitting. For correlated variables, principal components, which will transform a number of possibly correlated variables into a smaller number of uncorrelated variables, is another possible approach.

Also you mentioned that the standard deviation for the two estrogen variables, Estrone
and Bioavailable E2, are very large, and these measurements cannot be negative. This suggests these predictor variables might be highly skewed. We suggest plotting one of the outcome variables against one of the estrogen variables. If the data cloud is roughly football-shaped and there is a roughly linear relationship, no transformation is needed. Otherwise, you may want to consider taking the log or the square root of the estrogen variable, and plotting again to check whether there is a roughly linear trend. The reason is that linear regression models assume that the covariate and the dependent variable are linearly related. If some transformation of the covariate can make the plot roughly linear, the residual for the model will be smaller. In other words, the model will provide a better fit. In addition, we suggest plotting the histogram of the covariate to check whether the distribution of that variable appears at least roughly symmetric as this will tend to produce better-behaved residuals. However, transformation of the covariate to achieve symmetry could diminish the quality of the linear relationship between the covariate and the dependent variable.

Furthermore, the summary you provided shows that some of the covariates that were measured over three time points, such as SF36 Physical Health, do not appear to change much over time; at least the averages are roughly the same at all three time points. We suggest plotting these variables over time for individual patients to check whether they change a lot over time. If the changes are slight, we suggest using either just the baseline measurement or else the average of three measurements as a time independent covariate for each patient instead. However, it may be easier to interpret the model fitting result with the baseline measurement as a covariate instead of the average of the measurements at the three time points.
4.2 Assessment of Randomization

To verify whether the randomization of the trial was effective, we suggest comparing the eight background variables between the experimental and control groups. The background variables are Age, Education, Married or common law, Employment, Professional status, Caucasian ethnicity, Menopausal status, Body mass index. In addition to these eight variables, the baseline values of the other fifteen variables can be also used to compare the two groups. Two of these variables are continuous variables and the rest are sums of ordinal scales which I think that you are comfortable to treat as continuous variables.

For the continuous variables, Age, Education, Body mass index, and other fifteen variables at their baseline values, box plots are a convenient way to describe and visualize the data through its five-number summaries: minimum, lower quartile, median, upper quartile, and maximum. For each variable, side-by-side box plots for the two groups can be used to compare these variables visually; see Figure 1 for an example based on randomly generated data. Besides these plots, we suggest using two-sample t-tests to determine whether the two groups of patients can be viewed as being drawn from popu-
Table 2: Example of a two-by-two table

<table>
<thead>
<tr>
<th></th>
<th>Employment (Yes)</th>
<th>Employment (No)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control group</td>
<td>25</td>
<td>32</td>
</tr>
<tr>
<td>Experimental group</td>
<td>28</td>
<td>30</td>
</tr>
</tbody>
</table>

lations with the same mean by comparing the sample means (averages) of the two groups. The Wilcoxon rank sum test, a nonparametric alternative to the two-sample t-test to determine whether two samples come from different distributions, can also be applied.

For the binary variables, Married or common law, Employment, Professional status, Caucasian ethnicity, and Menopausal status, two-by-two tables can be used to examine the difference between the two groups; see Table 2 for an example of a two-by-two table. Besides these tables, Pearson’s chi-square test to test the equality of two binomial proportions or, in case the counts in any of the cells of the two-by-two table are small, the Fisher exact test, can be also applied.

4.3 Models to Predict Cognitive Function

First, we introduce a model for the outcome variable \textit{WMS auditory immediate} which includes only the most important covariate, \textit{Estrone}, as an example to introduce some basic ideas. Let \( i \) be the group indicator which takes on only two values, 0 or 1, where 0 means the patient belongs to the control group and 1 means the patient belongs to the experimental group. Let \( j \) be the patient number within the group, so that each patient is uniquely identified by the pair \( i \) and \( j \). Finally, let \( t \) be the time which takes value of 0 (baseline), 1 (6-month) and 2 (12-month). A simple model is:

\[
\text{WMS auditory immediate}_{ijt} = \mu_{it} + \beta \times \text{Estrone}_{ijt} + \epsilon_{ijt}.
\]
For the time being, the only thing we will assume about the error terms $\epsilon_{ijt}$ is that their expected values (means) are zero; in particular, because of the longitudinal nature of the data we need to allow for the possibility that the error terms corresponding to the repeated observations on the same patient are correlated. Taking the expectation on both sides of the equation, the model yields:

$$E(WMS \text{ auditory immediate}_{ijt}) = \mu_{it} + \beta \times Estrone_{ijt}.$$ 

This simple model uses Estrone to predict the expected value of WMS auditory immediate at different time points for the different groups and the different patients. The $\mu_{it}$ term in the model is to allow for a different overall level of response on WMS auditory immediate at each of three time points for each of the experimental and control groups. On the other hand, Estrone has subscripts $i$, $j$ and $t$, so the value of Estrone for each patient is allowed to be different at each time point.

Figure 2 illustrates an example for the expectation of WMS auditory immediate at the different time points for the two groups in the special case where is no relationship with Estrone ($\beta = 0$). This example is for illustration purposes only, and the actual figure could be totally different. The blue line stands for the control group and the green line stands for the experimental group; the dots on the lines represent the expectation of WMS auditory immediate at the different time points. The figure illustrates what might happen if the drug has a positive effect which means the expected value of WMS auditory immediate increases over time for the patients receiving the drug. Since the experimental group received the drug after the baseline measurement, the green line goes up from time 0 to time 6. The experimental group continues receiving the drug after
Figure 2: Mean of outcome variable at different time points (special case $\beta = 0$)

the 6 month measurement, so the green line goes up again from time 6 to time 12. For the control group, the patients only received the drug after the 6 months measurement. Therefore, the blue line is flatter up to 6 months but then goes up after 6 months.

It might be reasonable to assume the expectation of the outcome variable is the same in the two groups at baseline ($\mu_{00} = \mu_{10}$) since the patients were randomly assigned to the two groups. On the other hand, by specifying the model as indicated, the expectation is allowed to be different at baseline and the question of whether the data provide convincing evidence of such a difference ($\mu_{00} \neq \mu_{10}$) can be assessed as part of the analysis.

Perhaps the best way to look at this simple model when $\beta \neq 0$ is to take a group of patients who belong to a particular group and plot the implied regression line for the outcome versus Estrone using the data in that group at each specific time. Figure 3 illustrates this for the expectation of WMS auditory immediate against Estrone. The lines for the different times and groups are parallel as the slope $\beta$ in this simple model is
the same for all groups and times. The dots in Figure 2 are now the intercepts of these parallel regression lines (the values of the regression lines when Estrone = 0).

4.3.1 Treatment with Arzoxifen as a Moderator

Since you would like to use the treatment with Arzonxifen (Drug) as a moderator in the analyses, we introduce a new indicator variable, Drug, which takes on two values, 0 or 1, where the value 0 means the patient took the Drug in the preceding 6 months; and the value 1 means the patient did not. Each patient will have a value of Drug at each of the time points, and these may be different at different time points, so Drug also must have subscripts $i, j$ and $t$. Then, we expand the model by also incorporating the moderator variable, Drug. In general, moderator variables are variables that are assumed to modify (reduce or enhance) the influence of a specific independent variable on a dependent variable. To estimate the effect of a moderating variable in a linear
regression model, the model usually includes the independent variable, the moderator variable and an interaction between the independent variable and the moderator variable. Since, in this model, the time and group already determine whether the patient took the drug or not, the expanded model should only include the interaction term between Drug and Estrone, not the moderator variable, Drug, itself. The expanded model becomes:

\[
\text{WMS auditory immediate}_{ijt} = \mu_{it} + \beta_E \times \text{Estrone}_{ijt} + \beta_{DE} \times \text{Drug}_{ijt} \times \text{Estrone}_{ijt} + \epsilon_{ijt}.
\]

Again, if we take the expectation on both sides of the equation, the model yields:

\[
E(\text{WMS auditory immediate}_{ijt}) = \mu_{it} + \beta_E \times \text{Estrone}_{ijt} + \beta_{DE} \times \text{Drug}_{ijt} \times \text{Estrone}_{ijt},
\]

which is equivalent to:

\[
E(\text{WMS auditory immediate}_{ijt}) = \mu_{it} + (\beta_E + \beta_{DE} \times \text{Drug}_{ijt}) \times \text{Estrone}_{ijt}.
\]

The role of Drug as a moderating variable is assessed by estimating \( \beta_{DE} \), the regression coefficient for the interaction term, Drug \( \times \) Estrone; in particular, if \( \beta_{DE} = 0 \), then the model structure is as illustrated in Figure 3.

Figure 4 illustrates the expectation of WMS auditory immediate against Estrone when the moderator variable, Drug, has an effect (\( \beta_{DE} \neq 0 \)). The three blue lines corresponding to group and time combinations when Drug = 0 are parallel as these slopes are the same in this model: \( \beta_E + \beta_{DE} \times 0 = \beta_E \). The other three green lines corresponding to group and time combinations when Drug = 1 are also parallel but with
a different slope: $\beta_E + \beta_{DE} \times 1 = \beta_E + \beta_{DE}$. As in Figure 3, the dots in Figure 2 become the intercepts for the two sets of parallel regression lines in Figure 4. Additional covariates or moderator variables can be added into the model in a similar way.

### 4.3.2 Approaches for Fitting Longitudinal Models

To this point, we have focused on specifying a model for the expectation of the outcome variable. As already mentioned, there are repeated measurements of the outcome variables over time on each patient. The measurements for each patient are correlated, and the within-patient correlation needs to be addressed. There are two main approaches which can be applied to obtain valid inferences: linear mixed effects (LME) models and the generalized estimating equations (GEE) approach. It is often a good idea to analyze a dataset using different approaches to gain additional insights. Therefore we suggest using both LME and GEE models to analyze this longitudinal data.
LME models

LME models are linear models for longitudinal (or clustered) data that can be used to estimate the relationships between a continuous outcome and several predictors. These models typically include both fixed-effect parameters which describe the relationships between the covariates and the outcome variable, and random effects which are specific to patients within a population. The random effects are used to model the random variation in the dependent variable at different levels of the data, such as an outcome variable measured repeatedly over time on the same patient.

The model with only a random intercept and the model with both a random intercept and a random slope are two commonly used linear mixed effect models. We use these two simple models to illustrate some general ideas. For example, suppose we focus on the patients in one of the two groups in your data set; for simplicity, we suppress the subscript $i$. Let $\text{WMS auditory immediate}_{jt}$ be the outcome for patient $j$ at time $t$ in this group, and suppose the model includes only $\text{Estrone}$ as a covariate. The model with only a random intercept is:

$$\text{WMS auditory immediate}_{jt} = \beta_0 + b_0j + \beta_1 \times \text{Estrone}_{jt} + \epsilon_{jt};$$

here the fixed-effect parameters are $\beta_0$ and $\beta_1$ and the random intercept effect is $b_0j$. The random part of the model $(b_0j + \epsilon_{jt})$ is composed of a subject-specific random component $b_0j$ and an error term $\epsilon_{jt}$. In LME models, the random effect components are usually assumed to be independent of the error terms and also to be normally distributed with means of zero.

Random intercept models assume that each patient has a different intercept. Figure 5
Figure 5: Illustration of the simple linear model with random intercept

illustrates the model; for clarity, we draw only ten blue lines which represent ten different patients as an example. Each patient has only one random effect - a random intercept. In most cases, the line for each patient is not the primary interest. The red line in Figure 5, the regression line corresponding to a random intercept of 0 which represents the average of the individual lines, is usually the main interest. The model clearly can incorporate covariates such as Estrone with repeated measurements over time.

Expanding the model to also include a random slope leads to:

\[
\text{WMS auditory immediate}_{jt} = \beta_0 + b_{0j} + (\beta_1 + b_{1j}) \times \text{Estrone}_{jt} + \epsilon_{jt};
\]

here the random effects \(b_{0j}\) and \(b_{1j}\) are usually assumed to be independent of the error term \(\epsilon_{jt}\) and to have a bivariate normal distribution with means of zero. Figure 6 is an illustration of the model, where the ten blue lines stand for ten different patients and the
Figure 6: Illustration of the model with random slope and random intercept

red line is the average of these ten blue lines.

Random intercept models imply the correlations between pairs of measurements on
the same patient are the same for all possible pairs (exchangeable correlation structure).
On the other hand, random intercept and random slope models imply the correlations
between pairs of measurements on the same patient are different for different pairs since
the model allows both the intercept and slope to vary randomly across patients. Thus,
random intercept and random slope models have more flexibility in modeling the corre-
lation structure in the data.

Since LME models rely on distributional assumptions for the random effects and er-
ror term which are also assumed (usually) to be independent each other, they are quite
model-dependent. As an integrated part of the analysis, the assumptions underlying LME
models need to be checked, at least graphically, such as using residual plots to check the
normality of the error terms.
The GEE approach

Another possible approach for this project is to use the GEE approach. Instead of trying to model the within-patient covariance structure, this approach simply models the mean response. With this approach, the covariance structure does not need to be correctly specified to obtain consistent estimates of the regression coefficients and their standard errors. First, the regression model for the mean response is defined. The form of this regression model is totally flexible and it can be a simple linear model. Then a possible structure for the within-patient correlation, usually referred to as the “working correlation structure”, needs to be specified. The specified correlation structure needs not to be the true correlation structure, but is used in the algorithm that obtains the estimated regression coefficients and the estimated standard errors for the regression coefficients.

As an example, consider again the data for the patients in one of your groups. For simplicity, we suppress the subscript $i$. Let $\text{WMS auditory immediate}_{jt}$ be the outcome for patient $j$ at time $t$, and assume the model includes only Estrone as a covariate. Suppose the desired structure for the mean response is the simple linear model:

$$
\mu_{jt} = E(\text{WMS auditory immediate}_{jt}) = \beta_0 + \beta_1 \times \text{Estrone}_{jt}.
$$

Regardless of the choice of working correlation structure, the estimated regression coefficients obtained by the GEE approach are valid provided the sample size is large enough. A good choice of the working correlation structure, which is close to the true correlation structure of the data, will lead to more precise estimation of the regression coefficients. It is not easy to choose a good working correlation structure. Usually, the
working correlation structure is assumed to have a simple structure, such as independence, exchangeable, or AR1 structures. However, your dataset only has three (at most) repeated measurements for each patient, and the number of the patients is reasonably large \((n = 115)\), so we suggest using an unstructured working correlation structure which allows the correlation for each pair of within-patient measurements to be different.

The GEE approach provides “robust” standard errors which are obtained via a “sandwich” estimator. The robust standard errors are reasonable provided the sample size is large enough. More detail on the GEE approach can be found in Chapter 18 in the “Biostatistics - A Methodology for the Health Sciences” reference given at the end of the report.

### 4.3.3 Comparison of Approaches for Fitting Longitudinal Models

We have introduced two approaches, LME models and the GEE approach, which can be applied to obtain valid inferences for longitudinal data. However, there are several differences between these approaches.

First, your response variables are sums of ordinal variables, so they may not be normally distributed. LME models assume the random effects and error terms to be normally distributed with means of zero and that they are independent each other. However, the estimation in the GEE approach is based on the solution of generalized estimating equations that only rely on a specified model for the mean (and a working covariance structure). Distributional assumptions are not needed for the GEE approach.

Second, both LME models and the GEE approach provide estimated regression coefficients and standard errors. Inferences for the regression coefficients can then be carried
out in the usual fashion. Patient-specific inferences are possible for the LME models although these are usually not the primary interest. For example, in the random intercept model, the fixed-effect parameters $\beta_0$ and $\beta_1$ and the random intercept $b_{0j}$ for patient $j$ can be estimated; therefore the regression coefficients for each blue line in Figure 5 can be estimated. On the other hand, the GEE approaches are unable to make inferences for individual patients as these models are based only on the mean structure of the response.

In addition, model selection is an important part of the statistical analysis. For LME models, the usual likelihood-based model selection criteria can be used, such as likelihood ratio tests, AIC, etc. However, as the GEE approach does not depend on distributional assumptions, such model selection criteria are not available for the GEE approach. The only readily available model selection tool is Wald tests for the regression coefficients.

Finally, both LME models and the GEE approach can handle missing data. The GEE approach requires missing data to be missing completely at random (MCAR). That is, a measurement that is missing at a particular time point does not depend on either the observed or the unobserved data. On the other hand, LME models only require that missing data be missing at random (MAR) which is a weaker assumption. MAR allows the missing data mechanism to depend on both observed covariates and observed values of the dependent variable.

5 Summary

Several methods are suggested to compare variables at baseline to check whether the randomization of the trial was effective. For the continuous variables, box plots to compare the sample median visually, two-sample t-tests to compare the sample means, and
the Wilcoxon rank sum test to check whether two groups are from different distributions, can be applied. For the binary variables, two-by-two tables to show the proportions of Yes or No of each binary variable separated by the groups, Pearson’s chi-square test and Fisher exact test to test the equality of two binomial proportions, are suggested.

A model we suggest in Section 4.3 uses the most important covariate, Estrone, to predict cognitive function. An interaction term between Drug and Estrone is added to the model to meet the requirement that Drug should be a moderator. Additional covariates or moderator variables can be added into the model in a similar fashion.

Since the measurements for each patient are correlated, two main approaches which incorporate the within-patient correlation, LME models and GEE methods, can be used to obtain valid inferences. The GEE approach provides valid estimates of regression coefficients and their standard errors provided only that the model for the mean response is correctly specified and the sample size is large enough; no specific distributional assumptions are required. In contrast, the LME approach is quite model dependent. Therefore, although we suggest using both LME and GEE models to analyze this longitudinal data to gain additional insights, the GEE approach may be more appropriate than LME models for this project.

6 Reference