Statistical Analyses Plan for the Primary Efficacy Analyses in the DelpHi trial

Johannes Hertel*

March, 2016 (final editing November, 2016)

**Address for correspondence:*
Johannes Hertel
Department of Psychiatry, University Medicine Greifswald, Germany
Ellernholzstrasse 1-2, 17475 Greifswald
Tel.: ++49 – (0)3834/ 86 22 1 66; Fax.:++49 – (0)3834/86 68 89
e-mail: hertelj@uni-greifswald.de

1 Preliminary Remarks

The statistical analyses plan was written in its first version in November/Dezember 2015 and was then finalized in March 2016. The follow-up data was analyzed beginning April 2016 until October 2016. Thus, the analysis plan was written before analyzing the follow-up data. However, the baseline data was analyzed since 2013 and the design paper was published in 2011. At the time of the publication of the design paper, the study statistician (Johannes Hertel) was not yet employed in the DZNE and was therefore neither involved in designing the DelpHi-MV trial, nor in writing the design paper. The statistical analyses plan therefore is not coherent with the design paper in several points. The statistical analyses plan here only describes the efficacy analyses regarding the primary endpoints of the DelpHi trial.

The DelpHi cohort consists of individuals (above 70 years) who were screened positively on dementia. The sampling was done in primary care in the general practice (GP). Additionally, if possible, the principle care giver was included into the study. The GP was randomized by fair coin tossing to care as usual or intervention group. Thus, DelpHi is a cluster randomized clinical trial and all analyses have to respect the stochastic dependency of the data on the GP clustering. The randomization was done before baseline assessment of the individuals and the intervention cannot be classified as blinded, neither on the level of the GP, nor on the level of the study participant. The DelpHi trial can be classified as exploratory as no effect sizes a priori were known for the complex intervention that was performed. Moreover, the trial is clearly a pragmatic trial as it is implemented in primary care. Thus, the internal validity is potentially limited and several sources of bias (especially selection, attrition and performance bias) bias have to be assessed.

The power calculations mentioned in the design paper did not reflect the clustering of the data. The power of cluster randomized trials is not only a function of effect size, significance level and sample size, but dependent on the parameters of the concrete parameters of the clustering which were not known before sampling the data. In general, the power of a cluster randomized trial is lower in comparison to a randomized trial of equal sample size. The concrete loss of power is a function of the

intra-class correlation (ICC) of the outcome, the number and size of the clusters and the variability of the cluster sizes in the sample. It is known that with higher ICC the statistical power gets lower. If the ICC would be zero, normal power calculations apply. Thus, the sample size calculations noted in the design paper can be seen as upper bounds assuming that the ICC is zero for all outcome parameters. More accurate power calculations were therefore performed during the baseline sampling when first estimates of the ICCs where known and can be found at the end of this document.

2 Analysis Plan

2.1 Definition of the Primary Efficacy Endpoints

The design paper defined four dimensions of primary endpoints:

(1) Quality of life

(2) Caregiver burden

(3) Behavioural and psychological symptoms of dementia

(4) Pharmacotherapy with an antidementia drugs and prevention or suspension of potentially inadequate medication (PIM)

The fourth dimension includes two outcome measurements (antidementia drug treatment) and the suspension of PIM, thus, in total five outcome variables are derived. The first follow-up values on these dimensions (12 months after baseline) were defined as efficacy endpoints.

2.2 Operationalization of the Primary Endpoints

(1) Quality of life

Quality of life was measured using the Quality of Life in Alzheimer's Disease instrument [1]. It consists of 13 Lickert scale items that covers the following dimensions: physical health, energy level, mood, living situation, relationship with family members, caregivers and friends, memory as well as the ability to meaningful activities and financial situations. This instrument has been shown reliable and valid ratings by persons with mild to moderate dementia. Higher scores indicated better quality of life. The mean response on all answered items is used as outcome measurement, regardless whether the response on single items is missing. The mean response is treated as an interval scaled metric outcome.

(2) Caregiver burden

Caregiver burden was measured via the Berliner Inventar zur Angehörigenbelastung (BIZA-D) [2]. The BIZA-D was developed to assess objective and subjective burden due to caring for a PWD. It consists of 88 items covering 20 dimensions of caregiver burden. Objective burden is divided into six dimensions, assessed by 25 items: 1) basic care tasks, such as supporting eating and hygiene (seven items), 2) extended care tasks, such as supporting grocery shopping and legal affairs (three items), 3) motivation and guidance (four items), 4) emotional support (four items), 5) support of the maintenance of social contacts (3 items) and 6) supervision (four items). Each item has to be rated regarding the frequency of the support needed on a 5-point scale (e.g., supervision: Does the patient need this type of support 1=always, 2=mostly, 3=partly, 4=hardly, 5=not at all). Subjective burden is divided into a) the subjective burden of behavior change (six dimensions: burden due to cognition with four items, aggression with five items, depression with four items, disorientation with five items, late symptoms with three items, and loss of relationship with five items); b) the subjective burden of perceived conflicts between needs and responsibilities to care (six dimensions: burden due to personal constraints with nine items, negative evaluation of one's own caring with four items, missing social appreciation with four items, financial losses with four items, personal development with three items, missing institutional support with three items); and c) role conflicts (two dimensions; professional role conflicts with four items, family role conflicts with five items).

From analyzing the baseline data, it was known that the 20 sum scores of the scales were highly intercorrelated with each other. Consequently, in a principle component analysis, one principle component explaining 38.6% of variance dominated. As the sequential analyses of each scale would lead to massive multiple testing, the first principle component of the BIZA-D is chosen as outcome parameter, representing a compound measure of the caregiver burden. This compound measurement is treated as an interval scaled metric outcome.

(3) Behavioral and psychological symptoms of dementia

The behavioral and psychological symptoms of dementia were operationalized by the Neuropsychiatric Inventory (NPI), developed by the Alzheimer's Disease Cooperative Study investigators [3]. The NPI represents an interview by proxy on twelve dimensions of neuropsychiatric behaviors, i.e. delusions, hallucinations, agitation, dysphoria, anxiety, apathy, irritability, euphoria, disinhibition, aberrant motor behavior, night-time behavior disturbances, and appetite and eating abnormalities. The severity and frequency of each neuropsychiatric symptom are rated on the basis of scripted questions administered to the patient's caregiver. A total NPI score is calculated as the product sum of the frequency by severity scores within each domain. The NPI also assesses the amount of caregiver distress engendered by each of the neuropsychiatric disorders, but the caregiver distress scores were not used in our analysis. The total score of all twelve scales were used as interval scale metric outcome variable as supposed in the NPI manual.

(4) Pharmacotherapy with an antidementia drugs

Antidementia drug treatment was defined as in Wucherer et al. [4] The computer-based collection of primary data on medication in the context of the home medication review includes both prescription drugs (Rx) and over-the-counter (OTC) drugs. The assignment was integrated using a master file of the Pharmaceutical Index (GKV-Arzneimittelindex) [5]. Active substances were coded according to the Anatomical Therapeutic Chemical (ATC) classification system (German Modification)[6]. The outcome variable is dichotomous with individuals taking one of the following drugs (N06DA02: donepezil; N06DA03: rivastigmine; N06DA04: galantamine; N06DA52: donepezil and memantine; N06DX01: memantine) coded with "1". Individuals not taking one these drugs were coded with "0".

(5) Prevention or suspension of potentially inadequate medication (PIM)

Suspension of PIM taking was defined according to Priscus list [7]. A list of all ATC codes considered as PIMs is given in the appendix. From the baseline analyses, it was known that multiple prescriptions of PIMs are seldom in the study population (Wucherer et al. 2016). Thus, for efficacy analyses, we dichotomize the PIM prescriptions: "1"= one or more PIMs prescribed; "0"= no PIM prescribed.

2.3 Efficacy Analyses

2.3.1 Per protocol analyses

2.3.1.1 Definition of the analyses set

We define the per protocol analysis as complete case analyses. The only reason not to be included into the per-protocol analysis is due to missing data in the baseline assessment or the follow-up assessment. Contaminations by changing treatment assignments after randomization are impossible due to the study design. The main reason for missing data is supposed to be drop-out because of death or withdrawal of informed consent.

2.3.1.2 Primary efficacy Analyses

For primary efficacy analyses, for each of the main outcome of the study a separate generalized linear mixed model is fitted with the model specification corresponding to the scale level of the outcome under investigation. Thus, for quality of life, caregiver burden and neuropsychiatric symptoms linear mixed regressions will be used assuming a Gaussian distribution, while for antidementive drug treatment and PIM prescription logistic mixed models will be used. To account for the stochastic dependency of patients treated by the same GP, the GP will be included as random effect variable.

The follow up outcome variable at T_1 (twelve months after finishing the baseline) will be the dependent variable in these analyses; the study group is the predictor of interest, while the baseline value (T_0) of the outcome variable will be included as covariate to diminish residual variance and to account for inter-individual variance at baseline. Furthermore, sociodemographic variables (age, sex and living situation (alone vs. not alone)) are planned to be included as covariates, too. These variables can be supposed to contribute proportions of variance to the outcome variables which should be independent of possible study effects as the intervention cannot change age, sex or living situation. Thus, the inclusion of these variables should result in decreased residual variance enhancing thereby

the statistical power to detect differences between study groups. A positive intervention effect is then defined as a significant regression coefficient (one-sided test) of the study group variable. The p-values will be interpreted one-sided as one-sided hypotheses are tested. The design of the DelpHi trial allows for multiple testing (five primary outcomes), therefore, additionally the Bonferoni corrected p-values will be reported.

2.3.1.3 Sensitivity Analyses

Including the GP as random effect variable in the equation using random intercepts only implies that the effect of the GP on the outcome variable is independent of the effects of the predictors which may not be correct for the study group variable and the baseline variable of the outcome. Thus, to test the robustness of the potential treatment effect, in sensitivity analyses random slopes for the baseline variable of the outcome will be introduced into the models. Descriptively, these models allow for GP dependent treatment effect and thus will result in treatment estimates respecting the possibility of differentially collaborating GPs, thus assessing the robustness of the results regardless of a potentially present performance bias.

The treatment effect estimates from the different models will be compared descriptively against each other and, moreover, the random slopes models will be tested against the random intercept models via likelihood ratio tests. Note that these likelihood ratio tests are known to be overly conservative.

Furthermore, in the case of the metric outcomes, we will derive additionally confidence intervals and p-values via non-parametric bootstrapping using 2000 replications. These sensitivity analyses are planned to get confidence intervals and p-values independent on the parametric assumptions of Gaussian distributions and homoscedasticity. Especially in the case of neuropsychiatric symptoms, baseline analyses indicated a skewed distribution for the NPI scores.

2.3.1.4 Secondary Analyses

In secondary analyses, the above described statistical modelling will be performed stratified for the living situation (living alone vs. not living alone). As the intervention targets also the care-giver and

the social context of a person suffering from dementia, it is very plausible that the effects of the intervention are different for persons living alone and persons living not alone. For example, the care-giver burden might response differentially to the intervention. Thus, effect sizes of the intervention for the efficacy endpoints will be derived once for persons living alone and once for persons not living alone. Additionally, to allow an inference-statistical assessment, a study-group living status interaction term will be introduced in the models explained above. A significant interaction term will be interpreted correspondingly that the intervention does not have the same effect for individuals living alone and individuals living not alone.

Moreover, several subgroup analyses regarding differential treatment effects are planned following the workflow of the described stratification living alone vs. not living alone. The subgroups under consideration will be:

1) Moderate/severe vs. Mild dementia
2) Men vs. Women
3) Age under 80 vs Age above 80

2.3.2 Intention to Treat (ITT) Analyses

To control a possible attrition bias, intention to treat analyses will be performed using multiple imputation techniques for the imputation of missing data. The method of imputation was not defined in the first version of the analyses plan. It was defined in March 2016 after the study statistician had completed a workshop for the handling of missing data in the end of February in Stockholm (Metrika Consulting – The Nordic STATA Contributor).

2.3.2.1 Definition of the Analyses Set

The ITT analyses will be carried out on the subsample of baseline-sample with baseline data regarding the primary efficacy endpoint under consideration. Thus, individuals which dropped out during the baseline assessment before the efficacy endpoint was surveyed will be excluded from the intention to treat analyses as there is not the necessary data available to allow the imputation of the follow data.

Missing values in the follow-up variable of the primary efficacy outcome will be imputed if the baseline value is not missing. Hence the ITT analyses can be seen as partial ITT analyses. The attrition bias caused by drop-out between the measurement points will be controlled by the imputation. A possible attrition bias caused by drop-out during baseline assessment will be not controlled by the ITT analyses proposed here.

2.3.2.2 Methods of Imputation

The imputation model will be chosen congenial to the analyses model with the difference that the GP will be included as a fixed effect in the imputation model. Thus, the model of imputation will be a linear regression for the care-giver burden, quality of live and NPI and logistic for antidementive drug use and the PIM, including the respective baseline variable, age, sex, the living status and the study group variable as predictors. The method of imputation is chosen to be multiple imputations via chained equations, using 50 runs of imputations. Standardized effect sizes will be also computed from these imputations and compared to the estimates of the per protocol analyses. Note that an important statistical pre-requisite for the usage of imputations is that the missing of the outcome variable is stochastically independent from the value of the outcome variable at follow-up given its baseline value.

2.3.2.3 Primary Efficacy Analyses (ITT)

The ITT analyses will be performed following the schedule of the per protocol analyses with the only difference being that missing follow-up values will be imputed.

2.3.2.4 Sensitivity Analyses (ITT)

The sensitivity analyses described in the per protocol analyses are not feasible using multiple imputation for computational reasons (bootstrap, random slopes). We plan, however, to use a second imputation model (predictive mean matching) for sensitivity analyses to test the dependency of the results on the concrete imputation method. Thus, we will rerun the ITT analyses only with the

difference that the imputation models (including the same variables as before) will be based on predictive mean matching. Again, 50 runs of imputations will be used.

2.3.2.5 Secondary Analyses (ITT)

The same analyses will be carried out as described in the per protocol analyses with the difference that missing values in the outcome variable at follow-up will be imputed.

2.4 Methods against Bias

Here, we will describe the statistical treatment of potential bias. The methods used in the design of DelpHi-trial to reduce a priori certain kind of bias will be not discussed. Three types of bias need special attention in the DelpHi trial:

1) Selection Bias
2) Attrition Bias
3) Performance Bias

2.4.1 Selection Bias

Although the trial is randomized it is possible that there is selection bias leading to differential baseline values in the primary efficacy outcomes regarding control and intervention group. As the baseline value of the outcome variable is included as covariate in the primary analysis model, the effect sizes of the intervention are thereby independent of baseline differences. Regardless, we will screen the two study group on differences regarding primary and secondary outcome variables, sociodemographic variables and study parameters on differences using generalized linear mixed models with the baseline variable as dependent variable and the study group as predictor. The GP will be included as random effect. The model specification will be specified according the type of variable tested (e.g. logistic for binary variables, Gaussian for metric variables, poisson for count outcomes).

2.4.2 Attrition Bias

We will perform ITT analyses as described to reduce the impact of drop-out during the follow-up interval. To check, whether systematic drop-out during the baseline assessment may influence the results, we will run a drop-out analyses. To this point, we will fit multivariable logistic regressions with drop-out (yes/no) being the dichotomous outcome and the study group, sociodemographic and the screening value of the DEMTECT as predictors. These analyses will be performed three-times:

1) Drop-out overall
2) Drop-out due to death
3) Drop-out due to withdrawal of informed consent

If it would be possible to predict drop-out during baseline by study-group parameters, it would be possible that attrition bias is still present despite the ITT analyses. In this case, the results of the study have to be treated cautiously.

2.4.3 Performance Bias

As the DelpHi-Trial is conducted in the setting of primary care, the control of the treatment in the control group as "care as usual" is not possible. It could be that already the inclusion of a GP into the study will change the treatment of the included patient. This effect could easily mask potential intervention effects. Thus, a non-significant result cannot be easily interpreted. The treatment adherence of the GPs in the intervention group was evaluated and generally quite high. Nevertheless, it is very plausible that the GPs were different in their adherence to the study protocol. For this reason, we performed special sensitivity analyses (see section "per protocol analyses"). These sensitivity analyses deliver a weak test of the robustness of the effects despite the heterogeneity between the GPs.

2.5 Post Hoc Power Calculations

The actual power of the trial was assessed during the baseline sampling when first estimates of the ICCs where known, for reasons explained above. Table 1 gives the estimates for the ICC of the baseline values of the primary efficacy endpoints and the detectable intervention effect (power 80%) given an assumed coefficient of variation of cluster sizes of 1, a baseline follow-up correlation of 0.6,

a mean cluster size of 5 and 100 clusters (GP) included into analysis (total n=500). The power analyses were done with STATA 13/SE using the package *clustersampsi*. Note that the estimates reported in table 1 were derived while the baseline sampling was still ongoing.

Table 1: Detectable Difference (power 80%) between Intervention (n=250) and Control group (250)

	Intra-Class Correlation	Detectable Difference	Model
Quality of Life	0.061	0.26	linear
Neuropsychiatric Symptoms	0.361	0.43	linear
Caregiver Burden	0.192	0.35	linear

3 References

1. Logsdon, R.G., et al., *Quality of life in Alzheimer's disease: Patient and caregiver reports.* Journal of Mental health and Aging, 1999. **5**(1): p. 21-32.
2. Zank, S., C. Schacke, and B. Leipold, *Berliner Inventar zur Angehörigenbelastung - Demenz (BIZA-D).* Zeitschrift für Klinische Psychologie und Psychotherapie, 2006. **35**(4): p. 296-305.
3. Cummings, J.L., *The Neuropsychiatric Inventory: assessing psychopathology in dementia patients.* Neurology, 1997. **48**(5 Suppl 6): p. S10-S16.
4. Wucherer, D., et al., *Antidementia drug treatment in people screened positive for dementia in primary care.* J Alzheimers Dis, 2015. **44**(3): p. 1015-1021.
5. WIdO (Wissenschaftliches Institut der AOK) *GKV-Arzneimittelindex*. http://www.wido.de/amtl_atc-code.html, 2016.
6. (DIMDI), D.I.f.m.D.u.I. *Anatomisch-therapeutisch-chemische Klassifikation mit Tagesdosen. Amtliche Fassung des ATC-Index mit DDD-Angaben für Deutschland im Jahre 2014.* 2016 2016 [cited 2016 16.09.2016].
7. Holt, S., S. Schmiedl, and P.A. Thurmann, *Potentially inappropriate medications in the elderly: the PRISCUS list.* Dtsch Arztebl Int, 2010. **107**(31-32): p. 543-551.