

Statistical analysis plan – the Oslo Orthogeriatrics Study

Note: This statistical analysis plan was written prior to any unblinding of treatment allocation

1. Introduction

The aim of the Oslo Orthogeriatrics Study is to evaluate the effect of an orthogeriatric organization model for patients with acute fractures in the proximal femur. The detailed study protocol has been published (1), and the study is registered in ClinicalTrials.gov (NCT01009268). The aim of this document is, prior to any unblinding of the data, to establish details of the statistical analysis that were not finished when the protocol was published.

Initially, we recapitulate some of the most basic elements from the protocol.

Inclusion and exclusion criteria

Eligible patients were those admitted acutely to the Ullevaal Clinic of Oslo University Hospital, irrespective of age, for a femoral neck fracture, a trochanteric or a sub-trochanteric femoral fracture as result of a low energy trauma, defined as fall from own height or from a level not higher than 1 metre.

Patients were excluded if the hip fracture was part of multi-trauma or high energy trauma, if they were regarded as moribund at admittance (as judged by the admitting orthopaedic surgeon) and if they lacked a valid informed consent or assent.

Randomisation procedures

The randomisation was based on computer-generated random numbers, and carried out by a statistician (ES) without any contact with the patients or the personnel involved in the inclusion. We used block randomisation (blocks of variable and unknown size) to ensure an equal group size. The randomisation was stratified with respect to whether or not the patient was admitted from a nursing home, in order to get the groups balanced regarding pre-fracture cognitive decline, an important prognostic factor. The allocation of each patient (orthogeriatric or orthopaedic care) was by sealed, opaque, numbered envelopes that were held in the Emergency Department (different colours for the two stratification groups). For each randomised patient, the study physician (LOW) has checked that the randomisation envelope with the lowest number was used.

Consent and inclusion procedures were carried out by the orthopaedic surgeons on call in the emergency room.

Primary endpoint

A composite endpoint combining the Clinical Dementia Rating Scale (CDR) (2) and the 10 words memory task (immediate recall and the delayed recall parts) from the CERAD battery (3).

Secondary endpoints

- CDR and 10 words memory task analysed separately

- Preoperative and postoperative delirium, ascertained using the Confusion Assessment Method (CAM) (4).
- Duration and severity of delirium (according to the Memorial Delirium Assessment Scale - MDAS) (5).
- Incidence of dementia 12 months postoperatively (ICD-10-criteria for research), based upon consensus in an expert panel consisting of one experienced old age psychiatrist and one experienced geriatrician, who will have access to all clinical information and results of the cognitive tests, but are blinded to allocation.
- Results of other cognitive tests 4 and 12 months postoperatively (MMSE score, clock drawing test score).
- Length of hospital stay.
- Intra-hospital mortality and cumulative mortality at 4 and 12 months postoperatively. Causes of death will be ascertained from the Cause of Death Register, which is coded according to ICD-10.
- Residential status at 4 and 12 months and number of days in own home during the first four months, based upon best available information from patient, family, or the local municipality.
- The Barthel ADL Index score (6), adding up to a sumscore ranging from 0 to 20. The scoring is based upon proxy information from the best available source (relative, nurse etc). The Barthel Index is scored at 4 and 12 months postoperatively.
- The Nottingham Extended ADL Index (NEADL) (7), adding up to a sumscore of maximum 66. Scoring is based upon proxy information 4 and 12 months postoperatively.
- Score of the Short Physical Performance Battery (SPPB) (8) at 4 and 12 months.
- Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) (9). This is a proxy-based scoring system for cognitive decline over the last 10 years, and we have applied the original version at 12 months. At four months, we applied a modified version, asking for cognitive changes in the period since the operation, as previously described (10).
- The Cornell scale for depression in dementia (11) at 4 and 12 months.
- Persistent delirium at 4 and 12 months assessed with CAM.

Some of the secondary endpoints might be published in separate articles.

Surveillance variables

- Degree of mobilisation during the hospital stay, measured with the small body-worn accelerometer-based sensor ActivPal® (12).
- Nutritional indicators: Weight and weight changes, micronutrients in blood (Vitamin A, Thiamine, Pyridoxine, Folate, Vitamin B12, Vitamin C, 25-Hydroxy-Vitamin D, Vitamin E, Vitamin K, Homocystein, and Zink) four months after the fracture.

Blinding

All variables at the 4 and 12 months follow-up are collected by study nurses blind to allocation. The research personnel registering clinical data during the stay cannot be blinded, meaning in-hospital variables are not acquired blind to allocation.

Power calculation

No data were available allowing us to carry out precise power estimates based on our primary endpoint. Based upon previous experiences with CDR, however (13) (Engedal, unpublished

data), we judged 300 patients to be sufficient to detect clinically meaningful differences. As 20 % of hip fracture patients can be expected to die before the 4 months follow up (14), we aimed for 370 patients to be randomised. Due to constraints with respect to resources for further inclusion, we had to terminate inclusion when 332 patients had been randomised (1).

Patient flow

Patient flow is illustrated in Figure 1. Note that this is slightly different from the one published with the protocol (1), in that some minor misprints are corrected and number of patients at follow-up is added.

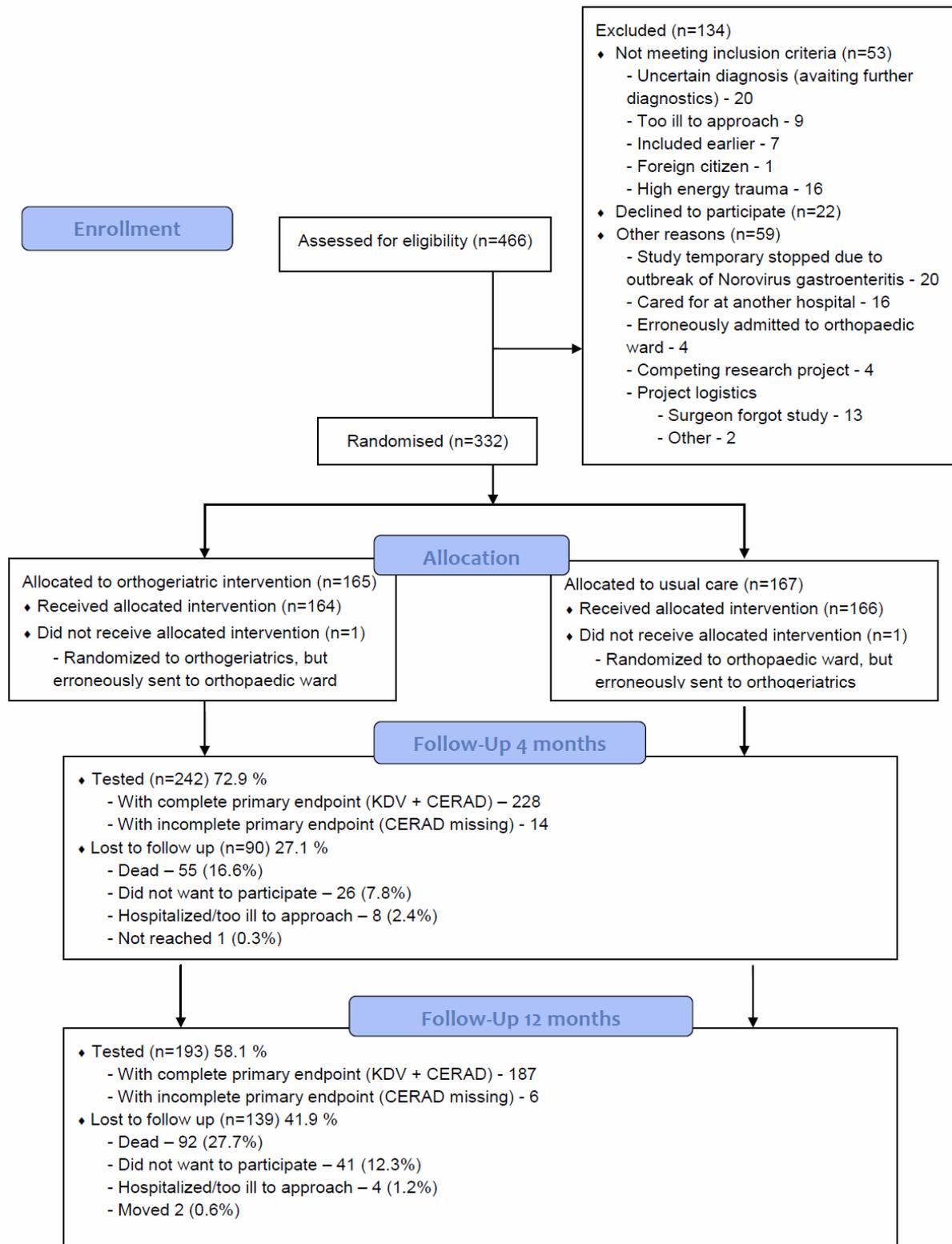


Figure 1. Patient flow after finishing the 12 months follow-up, March 2013

2. Statistical methods in general

In the protocol (1), it is stated:

"We will normalise these three scales [i.e. the CDR and the immediate and delayed recall parts of the CERAD 10 words memory test] into a 0–100 scoring (and reverse the CDR scoring as this is scaled in the opposite direction) to compute the composite endpoint. Before the randomisation allocation variable is added into the database, we will examine the proportion of patients in the lower vs the higher spectrum of performance as well as the correlation between each pair of scales. This information will be used to choose an appropriate weight to each of the scales before combining them. The distribution of the composite endpoint will also be examined."

And further:

"The primary analyses will be done in a modified intention-to-treat analysis including the sample of patients completing the CERAD 10 words test and CDR at 4 months. The weighting of the components of the primary endpoint (see primary endpoint section above) will be definitely decided upon before any analyses of the treatment effect are initiated. A separate per protocol analysis will be undertaken. The primary endpoint will be analysed by stratified linear regression. If the assumption of normality turns out to be violated, additional analyses will be performed on transformed data as well as by non-parametric methods (Mann–Whitney test). Binary outcomes will be analysed by chi square tests and logistic regression models. The material will be checked for any inequality in the distribution of important prognostic variables between the two arms, and if such are present, they will be adjusted for by including these in appropriate regression models.

Sensitivity analyses will be performed by imputing missing values in different conservative ways, e.g. by imputing the worst observed score of the treatment group and also by random sampling from the observed scores in the treatment group.

The main analyses of 12 months data will be carried out including patients with responses registered at that time point. Additional sensitivity analyses will be performed in the sample of all randomised patients using different methods of imputation as described above as well as by carrying the four month observation forward.

If an interaction test is statistically significant at the 0.10 level, we will analyse the effect upon the endpoints separately in the following subgroups: patients admitted from nursing homes or not, and patients with and without pre-fracture dementia.

We also register the time window from admittance to start of surgery, as any group difference here may also impact severely on the outcome."

In the further sections of this Statistical Analysis plan, we will elucidate these points

3. Studies of data distribution

During the autumn of 2012, we have scrutinized the distribution of the elements of the composite outcome variable as well as the composite itself, given different weighing principles for the elements. This was done blind to allocation and for the entire sample pooled. As expected, the two parts of the 10 words test were both skewed towards the left (floor effect), while the CDR (after transformation so that a higher score means a better function) was skewed to the right (ceiling effect).

We primarily studied the consequences of two weighing principles: a) The CDR, the immediate recall and the delayed recall weighs equally (1/3 on each), and b) the CDR weighs

50% and the two other 25% each. The distribution of the composite did not change much between these two weighing principles, but the latter seemed to give a slightly less prominent floor effect.

We have decided to use this 50/25/25 weighing of the three indices. We feel that this principle has the highest face validity, as the two parts of the 10 words memory test both assess aspects of verbal memory, while the CDR assesses a broader spectrum of dementia-related symptoms. The decision is also supported by the distribution assessments described above.

4. Primary effect analysis

The primary outcome (as explained in the above paragraph) will be analysed by linear regression with randomisation group as the explanatory variable (equivalent to a t-test). A two sided p-value below .05 will be taken as an indicator of statistical significance, and the mean difference in the primary outcome between the two randomisation groups with its 95% confidence interval will be reported. The sample for this analysis will be those who have fulfilled the 10 words test (both scales) and the CDR at the four months follow-up.

5. Handling of protocol violations

5.1 Inclusion of moribund patients

During the trial period, we became aware that some patients might erroneously have been included by the orthopaedic surgeon on call, even if they were obviously moribund at admittance. Patients that are likely to die before any intervention could be carried out, would attenuate possible intervention effects. Therefore, a clinician (TBW), blinded to group allocation, scrutinized the case notes of all patients who died during the primary stay (n=12). Three of these were considered to be moribund at admittance (all died before surgery), and thus erroneously included. They will be excluded before the primary analysis, but included in sensitivity analyses (see paragraph 6).

5.2 Patients not handled according to randomisation

Two patients were sent to the ward opposite to the one they were randomised to (see flowchart above). These patients will primarily be analysed according to the ward in which they were in reality treated, not according to the ward they were randomised to. A strict intention-to-treat analysis will, however, be carried out as a sensitivity analysis.

5.3 Patients for which a randomisation envelope was opened, but the patient was nevertheless not included

Such errors occurred on 13 occasions, due to the 24/7 inclusion procedure and the relatively high number of surgeons involved. Thus, 13 more envelopes have been opened than the number of included patients. The reasons are as follows:

- Included in the study before (previous fracture) - 6
- Randomised in a period where the geriatric ward was closed due to outbreak of gastroenteritis - 3
- Wrong diagnosis (patient had no hip fracture) - 2
- Patient sent to another hospital from the Emergency department - 1
- Initially opened envelope from the wrong stratum (see paragraph 1), error discovered immediately and new envelope opened - 1.

These patients were not included (except for the first six who already had been included earlier and the last one who was included based on the secondly drawn envelope), no data was registered from these hospital stays, and they are not further analysed.

6. Sensitivity analyses

5.1 Non-parametric analyses

Since the primary outcome is not normally distributed, the assumptions for parametric statistical methods may not be entirely fulfilled. We suppose that this will have only minor influence since the sample is relatively large, but as a sensitivity analysis we will also run non-parametric statistics (Mann-Whitney test of the primary outcome). Should the two analyses strategies lead to different conclusions, both will be reported.

5.2. Variables for adjustment

Variables with known or believed prognostic influence upon the outcome will be included in the regression model, one by one in addition to the randomisation group. If their introduction to the model changes the effect estimate (the mean difference, see paragraph 4 above) for the randomisation variable with 10% or more, they will be introduced in a final model including all variables with an effect of this size. The following variables will be subject to such analyses:

- Age
- Gender
- Number of years of education after primary school
- Admitted from nursing home or not (stratification variable)
- Score on the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE) (obtained in the acute phase)
- Preoperative waiting time
- APACHE II score (15)

5.3. Handling of missing data.

Missing values for the primary endpoint will be imputed in different ways in order to explore their potential influence on the results.

a) Patients who die before the 4 months follow-up

The primary analysis (section 4 above) will be carried out with deceased patients excluded. We will also report the effect of allocating the worst possible score for the primary outcome to these patients.

b) Patients missing for other reasons than death

- If a CDR score and the two scores of the 10 word memory test exist on that patient from the 12 months control, these values will be imputed to the 4 months dataset.
- If 12 months as well as 4 months data are lacking, we will carry out, and publish the results of three methods of data imputation:
 - The patients with complete data will be divided into four groups according to their pre fracture IQCODE score (quartiles), and for each patient with missing data, data will be imputed from a random patient from the same pre fracture IQCODE quartile.
 - Imputation of the worst score of the primary outcome registered from any of the patients with complete data.
 - Imputation of the mean score for the randomisation group the patient belongs to.

c) Protocol violations

We will also publish analyses with the three moribund patients included (see section 5.2), and with the two patients who were sent to the wrong ward analysed according to strict intention-to-treat (see section 5.3).

6. Analyses of secondary endpoints

Continuous variables will be analysed by two-sample t-test if the distribution is deemed approximately normal. Should there be obvious deviations from the assumption of normality the Mann-Whitney test will be applied. We will also apply the Mann-Whitney test for analysis of the clock drawing test, since this scale is ordinal with only six score levels. Categorical variables will be analysed by chi-square test.

7. Blind analysis

The project administrator (LOW) has been active in data acquisition, and might theoretically understand from the clinical data from a given patient who the patient is and whether the patient belongs to the control or the intervention group. The statistician (ES) and the principal investigator (TBW) have, however, had no role in data acquisition, and are completely blind regarding randomisation group. To prevent bias due to the project administrator's expectations, we will adhere to the following procedure: The project administrator (LOW), witnessed by a research assistant, allocates a random letter to patients in the intervention group and another to the patients in the control group. The code is written and stored safely, but not delivered to the statistician and the principal investigator. The statistician receives the dataset with these codes for the randomisation variable, and carries out the primary analyses. When the statistician and the principal investigator have agreed upon the results of the analysis, they will be unblinded.

8. References

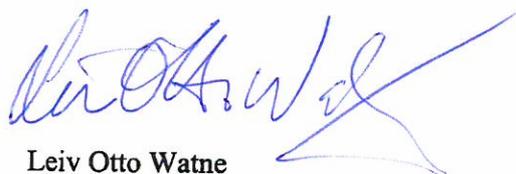
1. Wyller TB, Watne LO, Torbergsen A et al. The effect of a pre- and post-operative orthogeriatric service on cognitive function in patients with hip fracture. The protocol of the Oslo Orthogeriatrics Trial. *BMC Geriatrics* 2012; 12: 36.
2. Hughes CP, Berg L, Danziger WL et al. A new clinical scale for the staging of dementia. *Br J Psychiatry* 1982; 140: 566-72.
3. Morris JC, Heyman A, Mohs RC et al. The Consortium to Establish a Registry for Alzheimer's Disease (CERAD). Part I. Clinical and neuropsychological assessment of Alzheimer's disease. *Neurology* 1989; 39: 1159-65.
4. Inouye SK, van Dyck CH, Alessi CA et al. Clarifying confusion: the confusion assessment method. A new method for detection of delirium. *Ann Intern Med* 1990; 113: 941-8.
5. Breitbart W, Rosenfeld B, Roth A et al. The Memorial Delirium Assessment Scale. *J Pain Symptom Manage* 1997; 13: 128-37.
6. Mahoney FI, Barthel DW. Functional evaluation: the Barthel Index. *Md State Med J* 1965; 14: 61-5.
7. Lincoln NB, Gladman JRF. The extended activities of daily living scale: a further validation. *Disabil Rehabil* 1992; 14: 41-3.
8. Guralnik JM, Ferrucci L, Pieper CF et al. Lower extremity function and subsequent disability: consistency across studies, predictive models, and value of gait speed alone compared with the short physical performance battery. *J Gerontol A Biol Sci Med Sci* 2000; 55: M221-31.

9. Jorm AF. A short form of the Informant Questionnaire on Cognitive Decline in the Elderly (IQCODE): development and cross-validation. *Psychol Med* 1994; 24: 145-53.
10. Krogseth M, Wyller TB, Engedal K et al. Delirium is an important predictor of incident dementia among elderly hip fracture patients. *Dement Geriatr Cogn Disord* 2011; 31: 63-70.
11. Alexopoulos GS, Abrams RC, Young RC et al. Cornell Scale for Depression in Dementia. *Biol Psychiatry* 1988; 23: 271-84.
12. Grant PM, Ryan CG, Tigbe WW et al. The validation of a novel activity monitor in the measurement of posture and motion during everyday activities. *Br J Sports Med* 2006; 40: 992-7.
13. Barca ML, Engedal K, Laks J et al. A 12 months follow-up study of depression among nursing-home patients in Norway. *J Affect Disord* 2010; 120: 141-8.
14. Osnes EK, Lofthus CM, Meyer HE et al. Consequences of hip fracture on activities of daily life and residential needs. *Osteoporos Int* 2004; 15: 567-74.
15. Knaus WA, Draper EA, Wagner DP et al. APACHE II: a severity of disease classification system. *Crit Care Med* 1985; 13: 818-29.

9. Signatures

We hereby vouch for the fidelity of the study to this statistical analysis plan.

Oslo, Norway, 4th March 2013



Leiv Otto Watne
Project administrator
University of Oslo



Torgeir Bruun Wyller
Principal investigator
University of Oslo



Eva Skovlund
Biostatistician
National Institute of Public Health and University of Oslo