Statistical analysis plan for the trial:

Hypothermia in addition to decompressive hemicraniectomy in malignant MCA stroke: a randomized clinical trial

Hermann Neugebauer[1], MD; Hauke Schneider[2], MD, MBA; Julian Bösel[3], MD; Carsten Hobohm[4], MD; Sven Poli[5], MD; Rainer Kollmar[6], MD; Jan Sobesky[7], MD; Stefan Wolf[8], MD; Miriam Bauer[9]; Sascha Tittel[10]; Jan Beyersmann[10]; Johannes Woitzik[8], MD; Peter U. Heuschmann[11], MD; and Eric Jüttler[12], MD; for the IGNITE Study Group

[1] Department of Neurology, RKU – University and Rehabilitation Hospitals Ulm, University of Ulm, Germany
[2] Department of Neurology, University Hospital, Technische Universität Dresden, Germany; and Department of Neurology, Universitätsklinikum Augsburg, Augsburg, Germany
[3] Department of Neurology, University of Heidelberg, Heidelberg, Germany; and Department of Neurology, Klinikum Kassel, Germany
[4] Department of Neurology, University of Leipzig, Leipzig, Germany; and Klinikum Saalekreis, Merseburg, Germany
[5] Department of Neurology, University of Tübingen, Germany
[6] Department of Neurology, University of Erlangen, Erlangen, Germany; and Department of Neurology and Neurointensive Care, Klinikum Darmstadt, Germany
[7] Department of Neurology and Center for Stroke Research Berlin, Charité – Universitätsmedizin Berlin, Germany; and Department of Neurology, Johanna-Etienne-Krankenhaus Neuss, Neuss, Germany
[8] Department of Neurosurgery, Charité – Universitätsmedizin Berlin, Germany
[9] Center for Stroke Research Berlin, Charité – Universitätsmedizin Berlin, Germany
[10] Institute of Statistics, University of Ulm, Germany
[11] Institute for Clinical Epidemiology and Biometry, Comprehensive Heart Failure Centre, University of Würzburg; Clinical Trial Center, University Hospital Würzburg, Germany
[12] Department of Neurology, Kliniken Ostalb, Ostalb-Klinikum Aalen, Germany

Preface

The present trial was planned in 2010. Peter Heuschmann was consulted for methodology during the planning phase. Sascha Tittel conducted the statistical analysis under direction and supervision of Jan Beyersmann. The statistical analysis was approved as a master´s thesis from the University of Ulm, Faculty for Mathematics and Economics, Institute of Statistics.[1]

The preliminary statistical analysis plan was drafted during planning and beginning of the study and finalized after early stopping of the trial for safety reasons.

The first part of the present document is a summary of the statistical analysis plan. The second part of the document quotes verbatim (without quotation marks) from the master thesis considering defintions, data and study design, definition of data sets, statistical test and models and the statistical analysis.[1]

First Part: Summary of statistical analysis plan.

Introduction

In patients with space-occupying middle cerebral artery (MCA) infarction therapeutic hypothermia (TH) has been suggested additionally to decompressive hemicraniectomy (DHC). However, no sufficient evidence for the benefit of TH was available. Therefore, our objective was to conduct a trial to examine mortality and safety for patients that received TH in addition to DHC.

Study Methods

Trial design

The DEPTH-SOS trial is a multicentre, randomized controlled trial in six German academic centres to evaluate the effect of hypothermia (32-34°C, >72h) in addition to (DHC) (<48h) in adult MCA stroke patients (18 – 60 years) on mortality at day 14 after DHC (primary endpoint). Additionally, safety measures at day 14 and at 12 months, and functional parameters and mortality at 12 months are analyzed.

Randomization

Randomization is computer generated in blocks and stratified for centers using a web-based system (www.randomizer.at). Patients are assigned in a 1:1 ratio to either hemicraniectomy (control group) or hemicraniectomy plus therapeutic hypothermia (hypothermia group). The trial has an open-label design. Blinded rater obtaine follow-up-information after 12 months using a structured telephone interview. An independent institute of statistics analyzes the data.

Sample size

Sample size is to be 324 patients. For the calculation see the study protocol.[2]

Statistical interim analyses and stopping guidance

Safety analyses of all SAEs are planned after the inclusion of every 10th patient. Interim analysis of the primary endpoint is planned after treatment of 50 patients. Based on the results of the safety and interim analyses, an independent Data Safety and Monitoring Board that was not involved in the planning or conduction of the trial, recommends to continue or to stop the trial. For the interim analyses the concept of group-sequential tests for two proportions with two repeated significance tests is to be used. The alpha-spending function by Pocock is applied to be able to react quickly on possible differences in mortality between treatment groups.

Timing of final analysis

The final analysis is to be conducted when all patients have reached their respective 1-year follow-up.

Timing of outcome assessments

Mortality is measured when a patient dies, adverse events are measured when they occurre. Pneumonia is not accounted as SAE but as AE of special interest and is assessed after 14 days. The level

of consciousness on the National Institutes of Health Stroke Scale (NIHSS) is measured at hospitalization, before operation, and at hospital dismissal. The modified Rankin scale (mRS) score is measured at hospital dismissal and after 1 year for the follow-up analysis. The Glasgow Coma Scale (GCS) score is assessed after 14 days. The Barthel index is measured after 1 year. Retrospective consent to the respective treatment is assessed after 1 year from patients that are still alive, or relatives if the patient is not responsive or deceased.
Daily assessments include body temperature, intracranial pressure, CTs and MRIs y/n, pO2 value, medication y/n, osmotherapy y/n, and therapeutic ventilation y/n.

Statistical Principles

Confidence intervals and P values

The level of statistical significance is 5% for all analyses. There are no adjustments for multiple testing. Confidence intervals are at a 95% level, calculated via binomial distribution for dichotomous outcomes. Confidence intervals for ORs are calculated using Fisher's exact test. For recurring events and incidence rates log-transformed confidence intervals are calculated.

Analysis populations

The analysis is performed as intention-to-treat, with the crossover patients treated as having received TH. Additionally, a per-protocol analysis is performed for the outcome of mortality and SAEs after 14 days, excluding the crossover patients. For the same outcome as in the PP analysis, an as-treated analysis is performed additionally.

Trial Population

Eligibility

Patients with space-occupying MCA infarction were eligible for the trial, if they were to receive early DHC within 48h of symptom onset. Additionally, the following criteria had to be met:

- Age between 18 and 60,
- clinical sign of unilateral MCA infarction
- severe stroke, indicated by an NIHSS score of >=15 if the non-dominant hemisphere was affected or a score of >=20 if the dominant hemisphere was affected,
- impaired consciousness, indicated by NIHSS item 1a>=1,
- unilateral ischemia of at least 2/3 of the MCA territory, confirmed by CT or MRI; basal ganglia had to be at least partially involved,
- decision for DHC by treating physicians,
- possibility to begin DHC within 48h of symptom onset,
- possibility to start TH within 12h after DHC,
- written informed consent by patient or legal representative.

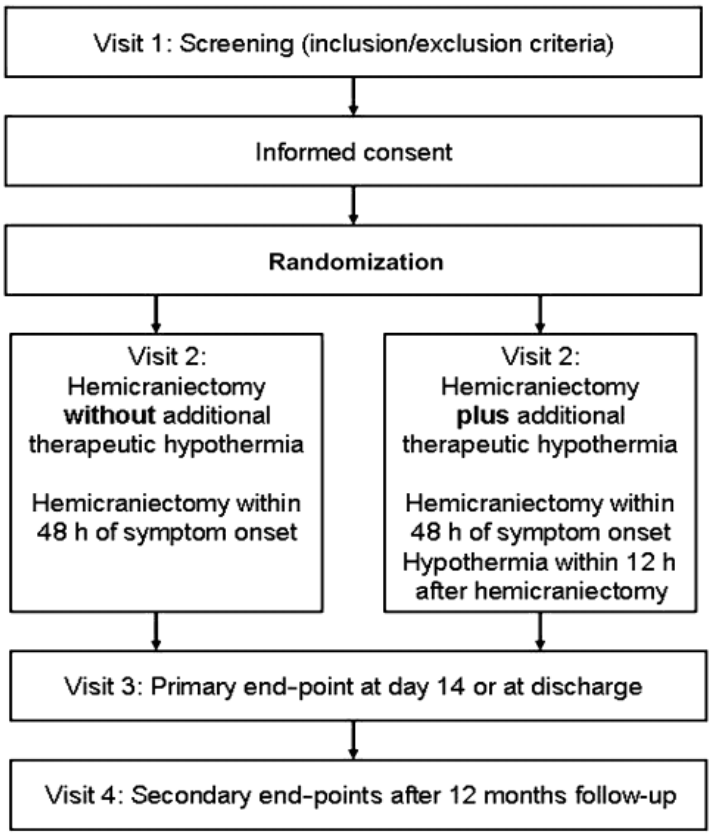
Exclusion criteria were any of the following:

- Premorbid mRS score>=2 and/or Barthel Index<95,

- simultaneous other brain lesions, e.g., traumatic brain injury, infarction (contralateral or infratentorial) in addition tot he index-infarction,
- clinical signs of transtentorial herniation,
- deep coma, indicated by a Glasgow Coma Scale (GCS) score<6 (does not apply to intubated patients),
- secondary hemorrhage in infarction area with space-occupying effect (PH2),
- known systemic bleeding or coagulation disorders
- known contraindication for TH, e.g., vasospastic disease, hematological disease with increased risk of thrombosis, paramyotonia congenita, severe pre-existing cardiac/liver/kidney disease,
- known indications for TH, e.g., after cardiopulmonary resuscitation,
- pregnancy,
- life expectancy of less than 3 years,
- sepsis,
- end-stage malignant disease,
- participation in other interventional trials.

Recruitment

Study flowchart:



For further recruitment information see study protocol.[1]

Withdrawal/follow-up

A blinded rater performs follow-up after 12 months using a structured telephone interview. Patients lost to follow-up are excluded from the analyses of the 12-month outcomes.

Baseline patient characteristics

Following characteristics of the study population are assessed at baseline:

- age in years (median, range)
- sex (no., %)
- mRS on admission (no., %)
- pre-existing Barthel Index on admission (median, range)
- the site of infarction (no., %)
- stroke in dominant hemisphere y/n (no., %)
- GCS (median, range; assessible in: no., %;)
- NIHSS total score on admission (median, range; assessible in: no., %;)
- time from onset of symptoms to randomization in hours (median, range)
- time from onset of symptoms to hemicraniectomy in hours (median, range)
- time from onset of symptoms to hypothermia in hours (median, range)
- adherence to assigned treatment (no., %)
- risk factors (arterial hypertension, diabetes, hyperlipidemia, present smoking, atrial fibrillation; no., %)
- rt-PA treatment (no., %)
- Onset to rt-PA i.v. in minutes (mean, standard deviation)
- Onset to rt-PA i.a. in minutes (mean, standard deviation)
- Onset to rt-PA m.R. in minutes (mean, standard deviation)

Analysis

Primery outcome

Primary (dichotomous) outcome is mortality after 14 days. The date of death is calculated as difference of day of death and day of admission in days.

Safety outcome

SAEs are classified as having at least one of the following:

- Blood and lymphatic system disorders
- Cardiac disorders
- Gastrointestinal disorders
- General disorders and administration site conditions
- Infections and infestations

- Injury, poisoning and procedural complications
- Investigations
- Metabolism and nutrition disorders
- Musculoskeletal and connective tissue disorders
- Nervous system disorders
- Psychiatric disorders
- Renal and urinary disorders
- Respiratory, thoracic and mediastinal disorders
- Skin and subcutaneous tissue disorders
- Surgical and medical procedures
- Vascular disorders

Body temperature is measured in °C and presented as mean per day. The intracranial pressure is measured in mmHg, presented as mean per day and classified critical if > 20 mmHg for a period longer than 10 minutes. Target temperature of TH was 33.0±1.0°C.
Pneumonia is rated as adverse event, not as SAE, because its rate in intubated ICU patients is reported to be 70% even under normothermia.
Scores on the modified Rankin scale range from 0 to 6, with 0 indicating no symptoms, 1 no substantial disability despite the presence of symptoms, 2 slight disability, 3 moderate disability necessitating some help, 4 moderately severe disability, and 5 severe disability; a score of 6 indicates death. Persons with a score of 0, 1, or 2 are considered to be functionally independent.
Scores on the Barthel index range from 0 (complete dependence) to 100 (independence) in increments of 5.
Scores on the Glasgow Coma Scale range from 3 to 15, with lower scores indicating reduced levels of consciousness.
Scores on the National Institutes of Health Stroke Scale (NIHSS) range from 0 to 42, with higher scores indicating more severe neurologic impairment.

Analysis methods

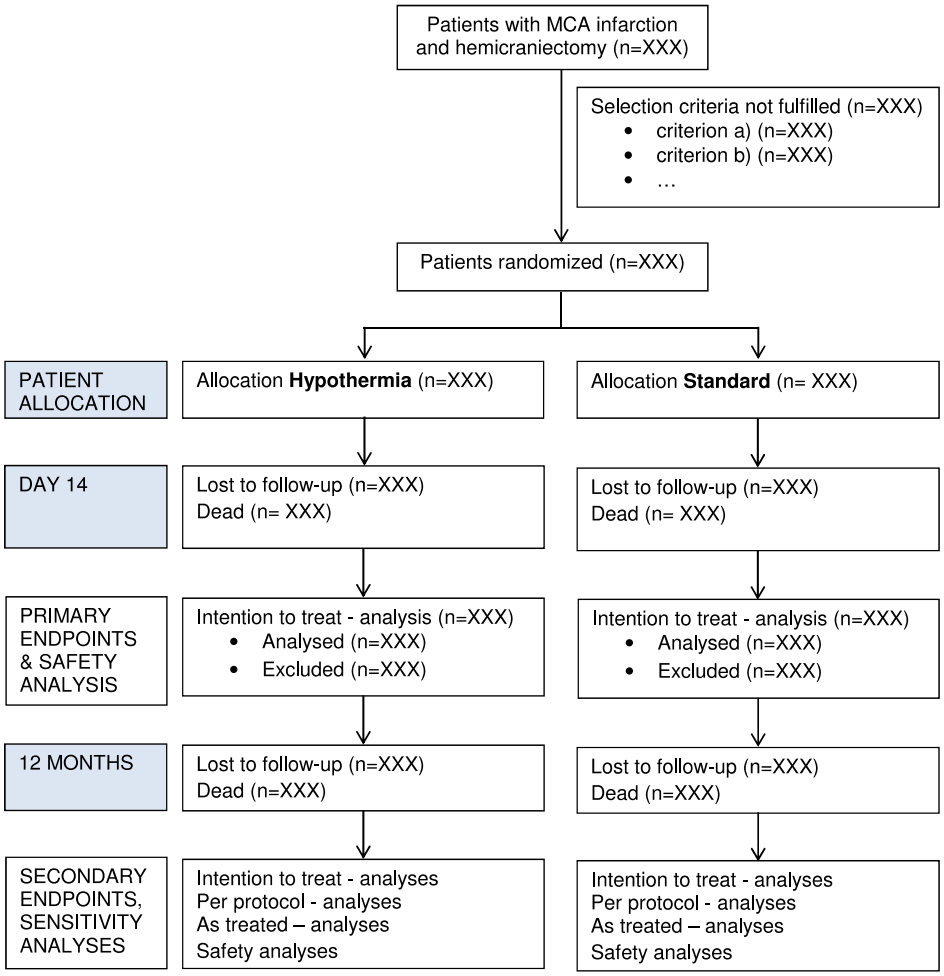
For mortality and safety endpoints after 14 days, binomial estimates with exact 95% CIs are calculated, as well as ORs with 95% CIs using Fisher's exact test. For sensitivity analysis, p-values of the χ^2-test are calculated additionally. For the safety outcomes we do not account for recurring events, but only considere the first event per patient. Pneumonia is analyzed analogously to the primary endpoint. Incidence rates for SAEs are calculated with 95% log-transformed CIs. The time from onset of symptoms to DHC is also log-transformed and compared with the t-test. Ventilation times are compared via Cox regression and log-rank test for the competing risk endpoints „end of ventilation" and „death under ventilation". The duration of intensive care treatment is also analyzed using log-rank test and Cox regression with combined enpoint „end of stay in intensive care" (dead or alive). For recurring events like days in intensive care with medication, number of CTs/MRIs, days with ventilation, days with osmo-

therapy, incidence rates with 95% log-transformed CIs are calculated, and Cox‘ proportional hazards model for competing risks (AE vs. Death w/o AE) are used. For functional outcomes NIHHS after 14 days, GCS after 14 days Wilcoxon’s rank sum test for group comparisons are used. A shift analysis is conducted for mRS scores after 12 months.

Missing data

Patients that are lost to follow-up are completely excluded from the analyses of the 12 month endpoints. Patients with missing data concerning functional scores are excluded from the respective analyses. Patients with missing data in incidence rate analysis for day 14 are censored at the first day without respective entry.

CONSORT flow-chart



Statistical software

Stastical analysis is performed using RStudio, Version 1.1.423.[4]

Trial registration

www.drks.de, Identifier DRKS00000623; URL:
https://www.drks.de/drks_web/navigate.do?navigationId=trial.HTML&TRIAL_ID=DRKS00000623b

Second Part: Statistical analysis plan - excerpts from the master´s thesis.[1]

Abbreviations

AE	Adverse Event
BI	Barthel Index
CI	Confidence Interval
CT	Computed Tomography
DEPTH-SOS	DEcompressive surgery Plus hypoTHermia in Space Occupying Stroke
DHC	Decompressive Hemicraniectomy
d14	Day 14 after inclusion into study (since randomization)
FU	Follow-Up
GCP	Good Clinical Practice
GCS	Glasgow Coma Scale
ICP	Intracranial Pressure
IQR	Interquartile Range
IR	Incidence Rate or Incidence Density Rate
ITT	Intention To Treat
log	Logarithm
MCA	Middle Cerebral Artery
MRI	Magnetic Resonance Imaging
mRS	modified Rankin Scale
NIHSS	National Institutes of Health Stroke Scale
OR	Odds Ratio
PP	Per Protocol
QOL	Quality Of Life
RCT	Randomized Controlled Trial
SAE	Serious Adverse Event
TH	Therapeutic Hypothermia
12mo	12 months after inclusion into study

Definitions

- Adverse Event
  By §3 (6) GCP-V: Any untoward event in a trial subject administered an investigational medicinal product and which does not necessarily have a causal relationship with this treatment.
- Serious Adverse Event
  By §3 (8) GCP-V: Untoward medical event that
  - is life-threatening or leads to death of a patient,
  - leads to or prolongs hospitalization,
  - causes permanent severe physical or psychological damage.
- Hazard

The hazard $\alpha_i(t)$ of the patient i regarding a specific event to the time point t is the conditional probability of the event happening within a very small time step, given prior survival and given that the patient is being observed just before t. Let $T \in (0, E)$ be an event time, where E denotes the end of observation. Then, the hazard of a patient is formally defined as follows:

$$\alpha_i(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(T \in [t, t + \Delta t) | T \geq t)}{\Delta t}.$$

- GCS

  The Glasgow Coma Scale (GCS) indicates the severity of an impairment of consciousness in response to defined stimuli. It is evaluated through three different tasks that a patient should accomplish: eye opening, verbal and motor response. The scale ranges from 3 (comatose or dead) to 15 (no impairment of consciousness).

- NIHSS

  The National Institutes of Health Stroke Scale (NIHSS) is a scale to evaluate the impairment caused by a stroke. The scale ranges from 0 to 42, where higher scores indicate more severe strokes. The max. score of 42 is assigned to deceased patients.

- mRS

  The modified Rankin Scale (mRS) is a tool used to measure disability after strokes. The score ranges from 0 (perfectly healthy) to 6 (death as a consequence of the stroke).

- BI

  The Barthel Index (BI) is a tool for the assessment of the autonomy of a patient, i.e., the ability to eat, walk and take care of personal hygiene autonomously. It does not assess the ability to live alone, because aspects like cooking, homemaking and social aspects are not considered. The scale ranges from 0 to 100, with higher scores relating to higher autonomy.

- p-value

  The p-value is interpreted as the conditional probability of an observation given that the null hypothesis is true.[4] Therefore the null hypothesis is dismissed at level $(1 - \alpha)$, if the p-value of the corresponding statistical test is less than or equal to α. This means, with a maximum probability of $\alpha \cdot 100\%$ we hold on to the null, even if it is false. Formally, for a test statistic T, the calculated value t of the test statistic and the null hypothesis H_0, the p-value is given by

  $$p = \mathbb{P}(T \geq t | H_0),$$

  which yields a connection to the significance level α. Let c be the the critical value, for which

  $$\alpha = \mathbb{P}(T \geq c | H_0)$$

  holds. Comparing both equations, it holds that

  $$p < \alpha \Leftrightarrow t > c.$$

  This means, the null is dismissed if and only if the calculated test statistic t is greater than the criti cal level c.

Data and Study Design

The DEPTH-SOS trial is a prospective, multicenter, open randomized controlled clinical trial (RCT) to evaluate the effectiveness of therapeutic hypothermia additionally to the standard therapy, which is the operative removal of a part of the skull, after a malignant middle cerebral artery infarction (most

severe, life-threatening stroke). The main question was, if additional TH could decrease mortality in stroke patients. The duration of the TH will be at least 72 hours with a target temperature of $33 \pm 1°C$, starting early within 12 hours after surgery.

According to sample size estimation and an expected absolute therapy effect of 12% $2 \cdot 162 = 324$ patients were included and assigned to the groups in a 1:1 ratio. The randomization of the patients has been executed through permuted blocks of 6 patients each. Patients had to be between 18 and 60 years to be included in the trial. Patients with a preexisting disability (score higher than 1 on the modified Rankin Scale (mRS)) or a preexisting impairment of daily activities (score below 95 on the Barthel Index (BI)) were exluded from the trial. Another exclusion criterion was a score of less than 6 on the Glasgow Coma Scale (GCS), indicating a deep coma.

From 2011 to 2015 50 patients (24 radomized to the control group, 26 randomized to the hypothermia group) from 6 centres throughout Germany were included. The study was advised to be stopped by the Data Safety and Monitoring Board in September 2015 for safety reasons. For the first interim analysis, planned after 50 patients, possible differences between treatment groups were to be indicated by Pocock's alpha-spending function. The interim analysis was cancelled due to the premature end of the study, instead the final analysis was conducted.

The primary endpoint was the mortality and safety analysis at d14 within the intention-to-treat (ITT) population. Since there were two crossover patients in the already small population, a per-protocol (PP) analysis of the primary and safety endpoints at day 14 was also conducted, with the data consisting of 48 patients (24 in control, 24 in hypothermia). Finally, an as-treated analysis was conduted, where the crossover patients were put in the control group for group sizes 26 (control) and 24 (hypothermia).

Methods

Outline

This chapter gives a comprehensive overview of the used methods. First the methods used for the d14 analyses are introduced along with the estimators followed by the statistical tests. Methods are introduced in the following order: Odds Ratio, incidence rate, Nelson-Aalen estimator, robust variance estimation, empirical cumulative distribution function along with the median and IQR, binomial estimates with the binomial test, confidence intervals, logistic regression, Fisher's exact test, the χ^2-test, the t-test, Wilcoxon's rank-sum test, and finally the log-rank test. For the analysis at 12mo and the other endpoints, sensitivity analyses for the NIHSS and GCS, and the mRS are introduced, and Cox' proportional hazards model and competing risks conclude this section.

Data Sets

Since in the randomized groups all known and unknown confounders are equally distributed, the focus was laid mostly on the ITT data set, i.e., the two crossover patients were included in our analyses and treated as if they were in their respective original group. Due to the small number of observations, deliberations about a possible bias from the crossover patients werde made, viz. if all SAEs would have occurred to them, whereas they did not get the hypothermia treatment. Therefore two additional analyses were conducted. In a second analysis, the PP approach was used for main mortality and

safety endpoints, where these two patients were excluded from the analysis to see whether the exclusion influences any outcome at all. Finally, the as-treated analysis was conducted where the two crossover patients were added in the control group.

The ITT data set has a size of 50 patients, with 26 being in the hypothermia group, and 24 in the control group. The PP data set has a total of 48 patients, with 24 in each group. The as-treated data set has 26 in control and 24 in hypothermia.

Odds Ratio

Since the data used for the Odds Ratio (OR) is binomially distributed, binomial estimates for the mortality after 14 days were calculated. The OR is a ratio of the odds of the occurrence of an outcome of interest between both groups. The OR ranges from 0 to 1, where an OR of 1 means that there is no difference between groups. The higher the value, the higher the odds for the occurrence of the outcome in the first group. Let p_e be the probability for a patient in the experimental group to have the outcome and p_c be the probability of a patient in the control group. The OR is then defined by

$$OR = \frac{p_e}{1 - p_e} \cdot \frac{1 - p_c}{p_c}.$$

When estimating the OR for confidence intervals, binomial estimates for p_e, p_c were used, instead. This calculation of the OR is valid for large populations. Given small numbers in the 2 by 2 table, the OR can have a great variance or even be biased. Therefore, another estimator was used that performs well and is used in the fisher.test function in R for the calculation of the OR: the Conditional Likelihood Estimator. See the section of Fisher's Exact Test for details.

Incidence Rate

The Incidence Rate (IR) is a measure for the occurrence of an event in a population. It estimates the hazard of of an event under the premise that the hazard is constant over time. The event may be one-time only, like death, or recurring, like medication. For example, the IR can be the sum of SAEs within the hypothermia or control group, divided by the sum of the patient-days during the observational period, i.e., from randomization to day 14 or to 12 months. The equation for the Incidence Rate of group $i \in \{1,2\}$ with group size $n_i \in \{24,26\}$ is

$$IR_i(t) = \frac{\sum_{j=1}^{n_i} N_{ij}(t)}{\sum_{j=1}^{n_i} d_{ij}(t)}.$$

Then $N_{ij}(t)$ is the number of events of the jth patient in group i to a predefined time point (14 days or 12 months), and $d_{ij}(t)$ denotes the patient-days of patient j in group i. A day with missing data was not added to the sum of events nor to the patient-days.

For recurring events like medication over the first 14 days the IR was used for group comparison. For the confidence intervals a log-transformation was performed to minimize data variability.

Nelson-Aalen Estimator

The Nelson-Aalen Estimator (NAE) is defined by

$$\hat{A}(t) = \sum_{u \le t} \frac{\Delta N(u)}{Y(u)},$$

where $\Delta N(u)$ is the number of oberved events in a population at time u and $Y(u)$ is the number of patients at risk at time u. The NAE is the sum of the quotient of the number of observed events at each time point and the population at risk at each time point. So the NAE estimates the cumulative hazard of a patient over time and is therefore used to compare with the observed IR. For the $(1-\alpha)$-CIs of the NAE it holds

$$\hat{A}(t) \mp z_{1-\frac{\alpha}{2}}\, \hat{\sigma}(t),$$

since the NAE is approximately normal, where

$$\hat{\sigma}^2(t) = \sum_{u \le t} \frac{\Delta N(u)}{Y^2(u)}$$

is a good estimator for the variance of the NAE.[5]

Robust Variance Estimation

Robust variance estimation was applied to the IRs, since the basic estimation of IRs relies on several assumptions, such as independent and identically distributed exponential event times, meaning that the hazard remains constant over time, which was assumed in the Incidence Rate section. In general, that is not plausible. That is why a robust estimator was used for the variance, which remains consistent even if the assumptions above do not hold true.[7]

Empirical Cumulative Distribution Function

The Empirical Cumulative Distribution Function (ECDF) is a descriptive function that assigns each value the proportion of values smaller than or equal to that value. The ECDF shows, e.g., the speed and rate at which patients reach the target temperature in the hypothermia group. The proportion of patients that reaches the target temperature at all can also be seen. The definition of the ECDF for a population of size n at time t is given by

$$F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{x_i \le t\},$$

where $x_i, i = 1, \dots, n$ is the time of an observation of interest of patient i and

$$\mathbf{1}\{x_i \le t\} = \begin{cases} 1, & x_i \le t \\ 0, & x_i > t \end{cases}$$

is the indicator function.

Median and Interquartile Range

The median m of a distribution is the value of a distribution function F for which holds

$$F(m) \ge \frac{1}{2} \quad \text{and} \lim_{t \to m} F(t) \le \frac{1}{2},$$

i.e., the median is the value m, for which the distribution function is at least 0.5 and the left-hand limit is smaller than or equal to 0.5. In the case of a non-continuous function like the ECDF, the first value m, for which the ECDF exceeds 0.5, is taken. In the case of patients reaching the target temperature, the median tells us when 50% of the patients have reached that target temperature. Median was chosen over mean, since the median is more robust in case the distribution is not symmetrical. The IQR is the range between the first and the third quartile of the distribution function, i. e., the 0.25% quantile and the 0.75% quantile.

Binomial Estimator and the Binomial Test

The binomial estimator is used when looking at proportions in a population or between populations. It is based on the binomial distribution with point probability

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x},$$

where p is the success probability in one trial, and n is the number of trials. In our case, p is the probability of an event in a population of size n. The estimator $\hat{p}$ is then given by the number of events divided by the size of the population.

For large n, the Central Limit Theorem can be used to approximate the binomial distribution by a normal distribution with mean np and variance $np(1-p)$. Let us denote with x the number of observed events and with p_0 a hypothetical success probability. Then the hypothesis $H_0: p = p_0$ is tested against the two-sided alternative $H_1 : p \neq p_0$ using the test statistic

$$T = \frac{x - np_0}{\sqrt{np_0(1-p_0)}},$$

which then is approximately standard normal distributed under the null. The null is rejected, if $T > z_{1-\frac{\alpha}{2}}$, where $z_{1-\frac{\alpha}{2}}$ denotes the $\left(1-\frac{\alpha}{2}\right)$-quantile of the standard normal distribution. Here, α is always 0.05, as is customary.

Confidence Intervals

A $(1-\alpha)$-Confidence Interval is an interval estimate. An interval for a parameter is said to be a $(1-\alpha)$-CI, if the probability of the parameter being within that interval is at least $(1-\alpha)\cdot 100\%$. Such intervals are obtained through assumptions on the distribution of a parameter. The test statistic for the binomial test is approximately standard normal. That means

$$\lim_{n\to\infty} \mathbb{P}\left(z_{\frac{\alpha}{2}} \leq T \leq z_{1-\frac{\alpha}{2}}\right) = 1-\alpha,$$

from which the asymptotic $(1-\alpha)$-CI for the parameter p can be derived:

$$\left[\hat{p} \mp z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right].$$

Due to the small sample size, exact CIs were used. The binom.test in R calculates exact CIs by Clopper and Pearson (1934) using two one-sided tests at $\left(\frac{\alpha}{2}\right)$ level. The lower and upper bounds (p_l, p_u) are obtained through the equations

$$\mathbb{P}(X \geq kp_l) = \frac{\alpha}{4} \quad \text{and} \quad \mathbb{P}(X \leq kp_u) = \frac{\alpha}{4}$$

respectively, where k denotes the number of successes.

Multinomial Logistic Model

For the analysis of the covariates ’age’, ’sex’, ’stroke severity’ and ’time from randomization to hemicraniectomy in hours’, a multinomial logistic model for ordinal and/or nominal variables was used. The variables here were ordinal with the exception of the nominal ’sex’. The goal was to estimate the probability of a patient to die within 14 days.

Denote with $\pi_i = \mathbb{P}(y_i = 1)$ the probability of patient i to die within the first 14 days, given his vector of p covariates x_i. Assume the odds of dying being expressed by

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha + x_i'\beta,$$

with $\beta = (\beta_1, \dots, \beta_p)'$, which gives us the log odds. α is the log odds of observing a death for a covariate vector $x_i = 0$ and the jth entry of the vector β is the log odds ratio of death per unit change in the jth covariate.

Due to the logarithmic output of the estimated coefficients and CIs in R, all coefficients and CIs were rescaled to the OR scale by taking the exponential of these values.

Fisher's Exact Test

Fisher's exact test is a method to test on independence between group assignment and the probability of a positive outcome. The null hypothesis can be written as $H_0 : OR = 1$ against $H_1 : OR \neq 1$, where OR is the Odds Ratio. Fisher's exact test was used for the primary and secondary endpoints, when there was a 2 by 2 table available with no null entries. The fisher.test in R calculates the p-value through the hypergeometric distribution in the 2 by 2 cases. For the two-sided test the pvalue is the sum of 'more extreme' entries in the table. That means, it simulates tables, where the entries are more extreme than in the observed table and sums up these probabilities to the p-value. The distribution function of the hypergeometric distribution is defined by

$$\mathbb{P}(X = k_1) = \frac{\binom{n_1}{k_1}\binom{n_2}{k_2}}{\binom{n_1+n_2}{k_1+k_2}},$$

where k_i is the number of events in group i and the group size n_i, for $i = 1,2$. Furthermore, denote by k_l and k_u the lowest and highest possible value for k_1. The p-value is then given by[6]

$$p = \sum_{x=k_l}^{k_u} \mathbf{1}\{\mathbb{P}(X = x) \leq \mathbb{P}(X = k_1)\} \cdot \mathbb{P}(X = x)\,.$$

For the calculation of the estimated OR, fisher.test uses not the unconditional Maximum Likelihood Estimator (MLE) introduced above, but the conditional MLE.

For the estimated OR and the limits of the CI the conditional noncentral hypergeometric likelihood is introduced,

$$L_c(k_1; OR) = \frac{\binom{n_1}{k_1}\binom{n_2}{k_2}(OR)^{k_1}}{\sum_{i=k_l}^{k_u}\binom{n_1}{i}\binom{n_2}{k_1+k_2-i}(OR)^{i}}.$$

From there, the estimator for the OR i derived via conditional MLE. Since there does not exist a closed form for the conditional MLE, one must use iterative procedures. The bounds of the CI for the estimated OR are then the values $\widehat{OR}_l$ and $\widehat{OR}_u$ that satisfy the equations

$$\frac{\alpha}{2} = \sum_{x=k_1}^{k_u} L_c\left(x; \widehat{OR}_l\right)$$

$$\frac{\alpha}{2} = \sum_{x=k_l}^{k_1} L_c\left(x; \widehat{OR}_u\right).$$

Due to the small sample size this exact test was used, the χ^2-test results were added as sensitivity analysis.

χ^2-test on Independence

The χ^2-test on Independence is usually used for large sample sizes, that are too big to evaluate with conventional methods. For example, the chisq.test in R gives a warning, if one cell of the 2 by 2 table used holds less than 5 observations.

The χ^2-test was used whenever Fisher's test was used on 2 by 2 tables, i.e., for both group and the other attribute of interest we have 2 levels. So the degree of freedom for the test statistic is $(2-1)\cdot(2-1)=1$. Let O_{ij} be the observed number in cell (i,j) of the 2 by 2 table. Denote with E_{ij} the expected frequency of cell (i,j); $i,j=1,2$ with total sample size n and n_i the total number of observations for attribute i, and n_j analogously defined. It holds

$$E_{ij}=\frac{n_i\cdot n_j}{n}.$$

The test statistic is then given by

$$X^2=\frac{\sum_{i,j=1}^{2}\left(O_{ij}-E_{ij}\right)^2}{E_{ij}},$$

which is approximately χ^2-distributed with 1 degree of freedom. Let $\widehat{X^2}$ be the value calculated with the data. Then the null hypothesis that the attributes are independent is dismissed at level α, if the p-value $\mathbb{P}\left(X^2>\widehat{X^2}\right)\leq\alpha$, i.e., the probabilty that the test statistic exceeds the value of the actually calculated value has to be smaller than the significance level.

t-test

The t-test is used to test for differences in mean values between groups. It uses the fact, that the sample means follow a t-distribution when centered and scaled properly. Denote with $x_{11},\dots,x_{1n_1}$ and $x_{21},\dots,x_{2n_2}$ the data in two groups. The underlying random variables are assumed to be normally distributed with (μ_1,σ_1^2) and (μ_2,σ_2^2) respectively. The test statistics is given by

$$t=\frac{\bar{x}_2-\bar{x}_1}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}},$$

which is approximately t-distributed with ν degrees of freedom, $\bar{x}_i$ denotes the arithmetic mean of group i and s_i the group standard deviation in group i.

The test statistic is also called Welch test statistic. It is also the default option of the t.test in R (which was used). It assumes no homogeneity in variances, which is regarded the safer variant instead of assuming same variances.[3] Then the denumerator in the test statistic is different and the degrees of freedom are integers. The number ν of degrees of freedom in the Welch test is calculated by

$$\nu=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{s_1^2}{n_1}\right)^2+\frac{1}{n_2-1}\left(\frac{s_2^2}{n_2}\right)^2},$$

which is the reason why the degrees of freedom are very often fractions instead of integers. The null hypothesis $H_0:\ \mu_1=\mu_2$ is then rejected, if $t<-t_{\nu,1-\frac{\alpha}{2}}$ or $t>t_{\nu,1-\frac{\alpha}{2}}$.

Wilcoxon's Rank-Sum Test

Wilcoxon's rank-sum test is an important non-parametric test based on the ranks of the observed data, not the data itself. So the underlying distribution does not have to be known. The test is used for hypotheses on homogeneity of two groups, i.e., if the underlying distributions of two groups are the same. The test assigns ranks to each observation. Define for the sample size $n = n_1 + n_2$ (the sum of both group sizes) and the data $X = (X_1, \dots, X_n)$ the rank of observation i as

$$R_i(X) = \sum_{j=1}^{n} \mathbf{1}\{X_j \leq X_i\}.$$

When ties occur in data, mid-ranks defined by

$$\tilde{R}_i(X) = \sum_{j=1}^{n} \frac{1}{2}\left(\mathbf{1}\{X_j \leq X_i\} + \mathbf{1}\{X_j < X_i\}\right)$$

are calculated. Wilcoxon's rank-sum statistic is given by

$$S_n = \sum_{i=1}^{n} \tilde{R}_i(X).$$

Note that if there are no ties in the data, it holds $R_i(X) = \tilde{R}_i(X)$. Wilcoxon's rank-sum test was used as sensitivity analysis for the two-sample t-test, e.g. the possible group difference in time from onset of symptoms to DHC.

Log-rank Test

The log-rank test is widely used to compare two survival curves. The test compares two groups at each death time and the expected number of deaths proportional to the population at risk at a specific time.[3] The result is then summed up over all death times and compared with the respective observed number of deaths.

Let n_1 and n_2 be the sizes of two groups we wish to compare at time zero. Denote with J the number of event times the first group, and $a_j, j = 1, \dots, J$ the number of observed events at time j in the same group. Let m_j be the total number of events in both groups at time j. Then the estimated expected number of events under the null hypothesis $H_0 : S_1(t) = S_2(t)$, i.e., the survival curves $S_1(t)$ and $S_2(t)$ are identical, is given by $n_{1j}m_j/N_j$, where $N_j = n_{1j} + n_{2j}$ is the total number of people at time j. The test statistic for the log-rank test is defined as

$$X_{LR}^2 = \frac{\left(\sum_{j=1}^{J} a_j - \frac{n_{1j}m_j}{N_j}\right)^2}{\sum_{j=1}^{J} \frac{n_{1j}n_{2j}m_j(N_j - m_j)}{N_j^2(N_j - 1)}},$$

which is approximately χ^2-distributed with 1 degree of freedom.[6] When there are no ties in the event times, it holds $m_j = 1$ for $j = 1, \dots, J$.

The log-rank test is non-parametric, meaning that there is no assumption about the specific distribution of the data, it is only assumed that both groups have the same survival function. The log-rank test was used each time where there were survival curves to compare. It also has a connection to Cox' proportional hazards model, where proportional hazards are assumed, indicating identical survival functions with the exception of a constant φ, i.e.

$$S_1(t) = S_2(t)^{\varphi}.$$

This is equivalent to

$$\exp\left(-A_1(t)\right) = \exp\left(-\varphi A_2(t)\right),$$

where $A_i(t)$ is the cumulative hazard function of group i. Differentiation of the exponents then yields

$$\alpha_1(t) = \varphi\alpha_2(t),$$

with $\alpha_i(t)$ denoting the hazard function of group i. This means the hazards of both groups are proportional to each other, which is the assumption of Cox' model. In the case of $\varphi = 1$, it yields the log-rank test, where possible differences in survival curves, and therefore hazards, are tested. The log-rank test is also connected to the IR, where ratios of occurred events and patient-days are calculated.

Analyses for NIHSS, GCS, BI and the mRS

For the NIHHS and GCS scores two different analyses were conducted, because there is no separate coding for deceased patients. In the NIHHS the scale ranges from 0 to 42, where higher scores mean more severe strokes. Therefore in the first analysis of the NIHSS score deceased patients were excluded. In a second analysis, the sensitivity analysis, the deceased patients were included and assigned a score of 42 to not condition on the future, i.e., no assumptions about the life status of patients were made.

The same approach for the GCS was made, which ranges from 3 to 15, where lower scores indicate more severe coma. In the first analysis of this score deceased patients were again excluded, and in the second the patients were included and assigned a score of 3.

The BI ranges from 0 to 100, where higher scores relate to higher autonomy. Only a descriptive analysis was conducted for this score due to missing data.

The mRS ranges from 0 to 6, where a higher score indicates a more severe disability following a stroke, with 6 being the score for death.

Cox' Proportional Hazards Model

Cox' proportional hazards model is used for regression of survival data similar to the usual linear or logistic regression. It also gives us the estimated hazard ratio between two groups, so we can see if there is significant difference in hazards and how strong the difference is. The Cox model assumes that both groups have the same baseline hazard $\alpha_0(t)$ at each time t. The baseline hazard, which is common to all patients, describes an existing hazard when the covariate vector x_i is zero, i.e., a hazard that is present even if no covariates influence the outcome. Then the conditional hazard at time t given the p-dimensional baseline covariate vector $x_i = \left(x_{i1}, \dots, x_{ip}\right)$ of patient i is given by

$$\alpha(t|x_i) = \alpha_0(t) \cdot \exp(\beta' \cdot x_i),$$

where $\beta = \left(\beta_1, \dots, \beta_p\right)'$ is the vector of regression coefficients and p is the number of covariates. This is called a semi-parametric model with $\beta \in \mathbb{R}^p$ being the parametric part, and the baseline hazard _0 the non-parametric part. The unknown baseline hazard cancels out when calculating hazard ratios. This model only holds true, if there are no competing risks, i.e., there is no other possible outcome than the outcome of interest. For example, when looking at the time of ventilation there are two possible endpoints, either 'death while ventilation' or 'end of ventilation without death', those two endpoints are the competing risks. Then there is more than one Cox model.

Let $\varepsilon \in \{1,2\}$ be a possible endpoint and let T be an observed event time, then the cause-specific hazard function is given by

$$\alpha_{0j}(t) = \lim_{\Delta t \to 0} \frac{\mathbb{P}(T \in [t, t+\Delta t), \varepsilon = j | T \geq t)}{\Delta t},$$

with $j \in \{1,2\}$.

The coxph function calculates the estimated logarithmic hazard ratio and the actual hazard ratio along with the confidence interval for the hazard ratio. Additionally, it calculates three statistical tests for the significance of the HR, the Wald test, the likelihood ratio test, and the score test, which is equivalent to the log-rank test.[3]

Statistical Analysis

Data

The DEPTH-SOS trial randomized 50 adult patients up to 60 years to either therapeutic hypothermia (26 patients, hypothermia group) or standard care (24 patients, control group). Since there were two crossover patients from the hypothermia group to the control group, an ITT approach was performed for the first analysis. In a second stept, a PP analysis and an as-treated analysis of mortality and safety were performed.

First of all a univariate analysis of the primary endpoint 'Mortality at d14' was performed. Therefore, only covariates from the baseline data (see Table 4.1) was used as well as for the multivariate analyses.

Endpoints

- Primary endpoint

  The primary endpoint is mortality at d14. It was analyzed dichotomously with a binomial estimate by group with exact 95% CIs. Exact 95% CIs for the OR using Fisher's test were also computed. Additionally, same method was used to test for independence between mortality and group membership. In addition, χ^2-tests were used as sensitivity analysis to verify the exact results. Then, multivariate analyses were conducted with baseline covariates 'age', 'sex', 'stroke severity', and 'time from randomization to hemicraniectomy'. The age variable was converted into an analyzed in decades, meaning the age of each patient was divided by 10. Our method here was logistic regression with four bivariate models, in which group membership was always included in addition to one of the baseline variables. Afterwards a model with group membership and all four baseline variables was set up. Outcome was again mortality at d14.

  Finally, a multivariate analysis was conducted, where all variables which fulfilled the AKAIKE criterion (univariate p-value ≤15.7%) were included. The method of choice was again logistic regression.

- Secondary endpoints at d14
  - Analysis of each (S)AE (P(Patient experiences (S)AE in [0,t])). This analysis did not account for possible recurring events, but only the respectively first (S)AE.
    - Evaluation analogous to primary endpoint, due to t=14d

- Calculation of the incidence rate $\frac{\sum_{j=1}^{n_i} N_{ij}(t)}{\sum_{j=1}^{n_i} d_{ij}(t)}$, with log-transformed 95% CIs by group
  - $N_{ij}(t)$: All (S)AEs of patient j in group i recurring at d14, and $i = 1{,}2$
  - dij(t): Total time patient j of group i spent in the study to a maximum of 14 days; total time ended with the minimum of death and t respectively.
  - n_i: Size of group i
  - Evaluation at d14 analogous to primary endpoint
- Respective time from onset of symptoms to DHC with log-transformation
  - t-test for comparison of groups
  - Wilcoxon's rank-sum test as sensitivity analysis
- Time from begin of TH to target temperature
  - Empirical distribution function (x-axis: time in days, y-axis: proportion of patients that reached target temperature)
  - Calculation of median and IQR
  - There was no comparison between groups, since there was no TH in the control group.
- Total time of TH
  - Empirical distribution function (x-axis: time in days, y-axis: proportion of patients that reached the end of TH)
  - Calculation of median and IQR
  - Again, there was no comparison between groups.
- Temperature load or adherence
  - Figures of the daily means over the first 14 days
  - Figures with minimum and maximum by group
- Total time of ventilation
  - Survival analysis (univariate score test (corresponds to the log-rank test) and Cox regression for all event specific risks (competing risks) with endpoints 'End of ventilation', 'Death while ventilation'), due to competing risks.
- Number of patients with tracheostomy at d14
  - Evaluation analogous to primary endpoint
- Total time of stay in intensive care
  - Survival analysis (univariate score-test and Cox regression for combined endpoint 'End of stay in intensive care' (alive+leave vs. Death in intensive care) and analysis of competing risks for the endpoints 'End of stay in intensive care alive', 'End stay in intensive care dead'), due to competing risks.

The following recurring events were analyzed with incidence rates and according to the following:

- 95% CIs based on log-transformation
- Graphical verification of goodness of fit with the Nelson-Aalen estimator
- Group comparison with Cox

- Log-rank test for group comparison
  - Medication while in intensive care
    - Incidence rate $\frac{\text{Number of daily doses of catecholamine}}{\text{Number of days (cumulated) patient stayed in intensive care}}$ by group
  - Usage of therapeutic ventilation
    - Incidence rate $\frac{\text{Number of days with ventilation}}{\text{Number of days (cumulated) patients stayed in intensive care}}$ by group
  - Number of CTs and MRIs while hospitalized
    - Incidence rate $\frac{\text{Number of CTs and MRIs}}{\text{Number of days (cumulated) patient spent hospitalized}}$ by group
  - Osmotherapy
    - Incidence rate $\frac{\text{Number of days with osmotherapy}}{\text{Number of days (cumulated) patient stayed in intensive care}}$ by group
- Secondary endpoints after 12 month FU
  - Incidence rate of SAEs $\frac{\sum_{j=1}^{n_i} N_{ij}(t)}{\sum_{j=1}^{n_i} d_{ij}(t)}$ with log-transformed 95% CIs by group
    - Evaluation after 12mo analogous at d14 and $d_{ij}(t)$ maximum 12 months
    - For all competing risks (AE vs. death without AE) Cox' proportional hazards model was used.
- Other endpoints
  - Stroke severity (NIHSS at d14)
    - Calculation of the mean was not appropriate due to the ordinal scale (0-42).
    - The higher the value, the more severe the stroke.
    - Calculation of the median, IQR and group comparison with Wilcoxon including only living patients
    - Sensitivity analysis, median, IQR with Wilcoxon; deceased patients were assigned a value of 42 (the maximum) to include all patients in the analysis. Furthermore, it was not conditioned on the future (that the patients were alive at d14).
  - Coma severity (GCS at d14)
    - Calculation of the mean was not appropriate due to the ordinal scale (3-15).
    - The lower the value, the more severe the coma.
    - Calculation of the median, IQR and group comparison with Wilcoxon including only living patients
    - Sensitivity analysis, median, IQR with Wilcoxon; deceased patients were assigned a value of 3 (the minimum) to include all patients in the analysis, analogously to the NIHSS.
  - Functional treatment results
    - The BI was analyzed only descriptively after 12mo due to incomplete data.
    - mRS
      - Ordinal scale, no mean
      - Categorized evaluation with 0-4 good, 5 bad, 6 deceased
      - Evaluation when leaving the hospital and after 12mo
      - Median, IQR, Wilcoxon group comparison

- Shift analysis after 12mo
  - Retrospective consent after 12mo
    - Dichotomous evaluation analogous to primary endpoint
  - Pneumonia at d14 (AE of special interest)
    - Analysis analogous to SAEs at d14
  - Graphical presentation of the ICP during the first 14 days per group
- PP analysis of SAEs and mortality
- AS-treated analysis of SAEs and mortality

References

[1] Tittel SR. *The Statistical Analysis of the Decompressive surgery Plus hypoTHermia for Space Occupying Stroke (DEPTH-SOS) Trial, with a View towards Mortality and Safety*. [master's thesis]. Ulm, Germany: Ulm University; 2018.

[2] Neugebauer H, Kollmar R, Niesen WD, et al; DEPTH-SOS Study Group; IGNITE Study Group. DEcompressive surgery Plus hypoTHermia for Space-Occupying Stroke (DEPTH-SOS): a protocol of a multicenter randomized controlled clinical trial and a literature review. *Int J Stroke.* 2013;8(5):383–387.

[3] Dalgaard P. *Introductory Statistics with R.* Second. University of Copenhagen, Denmark: Springer; 2008.

[4] Dorey F. The p-value: What is it and what does it tell you? *Clinical Orthopaedics and Related Research*. 2010;468(8):2297–2298.

[5] Klein JP. Small Sample Moments of Some Estimators of the Variance of the Kaplan-Meier and Nelson-Aalen Estimators. *Scandinavian Journal of Statistics.* 1991;18(4):333–340.

[6] Lachin JM. *Biostatistical Methods: The Assessment of Relative Risks*. Second. Hoboken, New Jersey: John Wiley & Sons, Inc.; 2011.

[7] Scosyrev E. Asymptotically robust variance estimation for person-time incidence rates. *Biom J*. 2016;58(3):474–488.