Moderne biostatistische Beiträge für Therapiestudien bei Schwindelsyndromen mit Tagebuch-basierten Attackendaten

Christine Adrion

2018

Aus dem
Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie (IBE)
der Ludwig-Maximilians-Universität München
Lehrstuhl für Biometrie und Bioinformatik
Direktor: Prof. Dr. rer. nat. Ulrich Mansmann

Moderne biostatistische Beiträge für Therapiestudien bei Schwindelsyndromen mit Tagebuch-basierten Attackendaten

Dissertation
zum Erwerb des Doktorgrades der Humanbiologie
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität München

vorgelegt von

Christine Adrion
aus Gräfelfing

2018

Mit Genehmigung der Medizinischen Fakultät
der Universität München

Berichterstatter:	Prof. Dr. rer. nat. Ulrich Mansmann
Mitberichterstatter:	Priv. Doz. Dr. med. Sandra Becker-Bense Priv. Doz. Dr. phil. Małgorzata Roos Prof. Dr. rer. nat. Christian Heumann
Mitbetreuung durch den promovierten Mitarbeiter:	—
Dekan:	Prof. Dr. med. dent. Reinhard Hickel
Tag der mündlichen Prüfung:	14.03.2018

Meinen Eltern in Dankbarkeit gewidmet.

"An approximate answer to the right question is worth a good deal more than an exact answer to an approximate problem."

—John W. Tukey (1915–2000)

Inhaltsverzeichnis

Abkürzungsverzeichnis

arcsinh	arcus-sinus-hyperbolicus
BDRM	Blinded Data Review Meeting
BEMED	MEnière's Disease with BEtahistine - Trial
CI	Confidence Interval
CONSORT	Consolidated Standards of Reporting Trials
CPO	Conditional Predictive Ordinate
DIC	Deviance Information Criterion
DSGZ	Deutsches Schwindel- und Gleichgewichtszentrum
EA2	Episodische Ataxie Typ 2
EMA	European Medicines Agency
FAS	Full Analysis Set
FDA	Food and Drug Administration
GLMM	Generalized Linear Mixed Model
ICH	International Conference on Harmonisation
INLA	Integrated Nested Laplace Approximation
ITT	Intention-to-treat
KI	Konfidenzintervall
M(C)AR	Missing (Completely) At Random
MCMC	Markov Chain Monte Carlo
NB	Negativ Binomial-Verteilung
PIT	Probability Integral Transform
Poi	Poisson-Verteilung
PP	Per-Protocol
PRO	Patient-Reported Outcome
QoL	Quality of Life
R	R : A language and environment for statistical computing
RCT	Randomized Controlled Trial
RR	Rate Ratio
SAP	Statistical Analysis Plan
SAS	Statistical Analysis System
SOP	Standard Operating Procedure
SPIRIT	Standard Protocol Items: Recommendations for Interventional Trials
VDADL	Vestibular Disorder Activities of Daily Living Score
ZI	Zero-Inflation

Zusammenfassung

Hintergrund: Morbus Menière ist eine chronisch progrediente Erkrankung des Innenohrs, gekennzeichnet durch anfallsartig auftretende Schwindelepisoden mit Hörminderung, Tinnitus oder Druckgefühl im betroffenen Ohr. Als first-line symptomatische Therapie zur Prophylaxe von Schwindelattacken bzw. Reduzierung der Attackeninzidenz gilt eine längerfristige, medikamentöse Behandlung mit dem Wirkstoff Betahistin-dihydrochlorid, zugelassen in der maximalen Tagesdosis von 48 mg. Es existiert keine ausreichende Evidenz hinsichtlich der Wirksamkeit der etablierten Betahistin-Therapie, v. a. aufgrund methodischer Mängel bisheriger randomisierter kontrollierter klinischer Studien (RCTs). Ziel dieser Arbeit ist die Beschreibung und Anwendung methodischer Verfahren und biometrischer Prinzipien bei der Entwicklung eines Statistischen Analyseplans (SAP) für eine verblindete konfirmatorische Phase III RCT mit longitudinalen Zähldaten am Beispiel der BEMED (MEnière's Disease with BEtahistine)-Studie (Parallelgruppendesign; 3-fache Dosis vs. Standarddosis vs. Placebo).

Methoden: Die Wirksamkeitsdaten (Attackenrate pro Zeiteinheit) wurden abgeleitet anhand von patientenberichteten Rohdaten in papierbasierten Schwindel-Tagebüchern, welche ein unverzichtbares Instrument zur kontinuierlichen Dokumentation des individuellen Attackenstatus und zur Erfassung von patientenorientierten Efficacy-Endpunkten bei RCTs zu symptomatischen chronischen Erkrankungen darstellen. Die Primäranalyse nach dem Intention-to-treat Prinzip erfolgte modellbasiert: Ein Generalized Linear Mixed Model (GLMM) unter Annahme der Negativ Binomialverteilung berücksichtigt den patientenindividuellen kompletten Verlauf der Attackeninzidenz sowie die Anzahl an Beobachtungstagen pro Zeiteinheit bei Annahme eines Missing-at-Random Mechanismus. Diese Analysestrategie erlaubt einen adäquaten Umgang mit unterschiedlichen Dropout-Situationen und komplexer Missingness-Struktur aufgrund unvollständiger Tagebuch-Dokumentation.
Auf der Basis einer vergleichbaren offenen Vorstudie und über simulierte Daten erfolgte eine vorhersageorientierte Selektion und Validierung alternativer Modellspezifikationen. Hierbei lag der Fokus auf prüfbaren, typischen Annahmen in der Zähldatensituation, insbesondere bezüglich (i) der zugrundeliegenden Verteilung (Poisson, Negativ Binomial, Modifikationen für Zero-Inflation, varianzsstabilisierende Transformation), (ii) der Random Effects-Struktur, (iii) des Response-Profils über die Zeit (Mittelwertstruktur). Im SAP vorab festgelegte Bayesianische Tools wie das DIC, Leave-one-out kreuzvalidierte Kriterien basierend auf der posteriori prädiktiven Verteilung der Daten für den Hauptendpunkt, oder Bewertungsregeln (Proper Scoring Rules) wie der Log-Score zur Beurteilung der Güteeigenschaften der prädiktiven Verteilung (Kalibrierung und Trennschärfe), ermöglichen eine informierte Entscheidung für ein adäquates GLMM für die Primäranalyse der BEMED-Daten.

Ergebnis und Schlussfolgerungen: Wirksamkeitsanalysen bei verblindeten konfirmatorischen RCTs mit longitudinalen Zähldaten und papierbasiertem Patiententagebuch (Symptom-Kalender) bedürfen einer komplexen Methodik, sowohl bei der Studiendurchführung inklusive der adäquaten und möglichst objektiven Ableitung der Efficacy-Daten,

als auch bei der Präspezifikation einer validen und robusten Analysestrategie für den SAP. Im Vergleich zu frequentistischen Ansätzen existiert im Bayesianischen Setting eine mächtige Toolbox zur Evaluation der prädiktiven Performance konkurrierender gemischter Modelle. Zu wenig Beachtung in der klinischen Forschung zu Schwindelsyndromen findet bislang die Entwicklung und Validierung von krankheitsspezifischen Patiententagebüchern, welche im kontrollierten Setting verblindeter Phase III Therapiestudien mit pragmatischer Fragestellung (Effectiveness) zum Einsatz kommen.

Summary

Background: Menière's disease is a chronic progressive disorder originated in the inner ear characterized by devastating vertigo spells with hearing loss, tinnitus or aural fullness in the affected ear. The first-line therapy to prevent or to reduce the incidence of Menière's induced vertigo episodes is a long-term prophylactic treatment with betahistine-dihydrochloride approved in a dosage of 48 mg daily. There is limited evidence to support the effectiveness of betahistine therapy mainly due to low quality randomized controlled trials (RCTs) or observational studies. The objective of this thesis is to describe and apply the key methodology and statistical principles needed to develop the statistical analysis plan (SAP) for a blinded confirmatory phase III RCT with longitudinal count data as primary efficacy outcome. Our approach is illustrated by the large scale BEMED (MEnière's Disease with BEtahistine) trial comparing the efficacy of two different doses of betahistine versus placebo treatment.

Methods: Efficacy data (number of attacks per time unit) were derived from patient-reported raw daily data collected by paper-based vertigo diaries. In RCTs of symptomatic chronic diseases this instrument is essential to continuously record the patients' subjective experience of disease events (vertigo attacks) and to provide patient-centered efficacy endpoints. A model-based primary analysis consistent with the intention-to-treat principle was conducted: A Generalized Linear Mixed Model (GLMM) with a negative binomial distribution considers the entire patient-specific profile of attack frequencies over time together with the number of evaluated days per time unit assuming a missing at random mechanism. This analysis strategy enables to adequately handle different types of dropouts and complex missing data situations.
On the basis of a comparable open-label study conducted in advance of the definitive RCT and by means of a simulation study a prediction-oriented selection and validation of rival model specifications was performed. We focused on common testable assumptions for count response data, particularly in terms of the (i) underlying distribution (Poisson, negative binomial, modifications for zero-inflation, variance-stabilizing transformation), (ii) random effects structure, (iii) response profile over time (mean structure). A priori specified Bayesian tools for model criticism such as the DIC, leave-one-out cross-validated criteria based on the posterior predictive distribution of the data for the primary outcome, and proper scoring rules (e.g. the logarithmic score) to evaluate and compare the predictive capability of different competing models were applied to prepare the SAP. These techniques facilitate an informed decision making for choosing an adequate GLMM for the pre-planned principal analysis of the BEMED trial before the blind is broken.

Results and Conclusions: In double blind confirmatory RCTs for vestibular syndromes with longitudinal count data obtained from paper-based patient diaries (symptom calendars) a complex methodology is needed within the scope of efficacy analyses, starting at the planning stage, the trial conduct which involves rigorous rules for the derivation of the efficacy outcomes, up to the pre-specification of a valid and robust analysis strategy required for the SAP. In contrast to frequentist approaches a powerful Bayesian toolbox is

available in order to evaluate the predictive performance of competing mixed effects models. The development and validation of disease-specific patient diaries implemented in the regulatory setting of blinded phase III pragmatic trials (designed to measure effectiveness) is of paramount importance in order to produce sensitive, meaningful and interpretable trial endpoints.

1 Einleitung

1.1 Klinischer Hintergrund und Motivation

1.1.1 Allgemeines zu Schwindelerkrankungen

Als Schwindel bezeichnet man entweder eine unangenehme Störung der räumlichen Orientierung, oder die fälschliche Wahrnehmung einer Bewegung des eigenen Körpers, das heißt Drehen und Schwanken, und oder der Umgebung. Schwindel ist keine Krankheitseinheit, sondern das Leitsymptom verschiedener Erkrankungen unterschiedlicher Ätiologie und Pathogenese, welche vom Innenohr (das Labyrinth oder den Nervus vestibularis betreffend) oder vom Hirnstamm oder Kleinhirn ausgehen, aber auch psychische Ursachen haben können (Strupp & Brandt, 2008; Strupp *et al.*, 2013). Nach Brandt *et al.* (2004) stellen peripher vestibuläre, zentral vestibuläre sowie somatoforme Schwindelformen neben Kopfschmerz das häufigste Leitsymptom in der Neurologie dar. Die Lebenszeitprävalenz von Dreh- oder Schwankschwindel liegt bei etwa 20 bis 30%, mit einer erhöhten Schwindelprävalenz im Alter, die 12-Monats-Inzidenz für Schwindel insgesamt beträgt ca. 3% (Neuhauser, 2007, 2009).
In der Schwindelambulanz des Deutschen Schwindel- und Gleichgewichtszentrums (DSGZ) am Klinikum der Universität München sind die häufigsten Diagnosen der benigne periphere paroxysmale Lagerungsschwindel (BPPV) mit etwa 17.1%, der phobische Schwankschwindel mit 15.0%, und zentral vestibuläre Schwindelsyndrome (12.3%). Die vestibuläre Migräne stellt mit 11.4% die häufigste Ursache spontan rezidivierender Schwindelattacken dar (Strupp *et al.*, 2013). Häufige Diagnosen sind zudem Morbus Menière (10.1%) und Neurititis vestibularis (8.3%). Weitere seltenere Schwindelerkrankungen sind die Vestibularisparoxysmie (3.7%), gekennzeichnet durch kurze heftige, Sekunden bis wenige Minuten anhaltende Dreh- oder Schwankschwindelattacken mit oder ohne Ohrsymptome wie Hörminderung und Tinnitus (Hüfner *et al.*, 2008), sowie die episodische Ataxie Typ 2 (EA2), charakterisiert unter anderem durch rezidivierende, meist Stunden bis Tage anhaltende Attacken mit Schwindel und Gang-, Stand- oder Extremitäten-Ataxie und zentralen Okulomotorikstörungen (Brandt *et al.*, 2004; Strupp *et al.*, 2008b).

Im allgemeinen werden Schwindelsyndrome klassifiziert nach 1.) der *Art* des Schwindels (Drehschwindel (wie Karussellfahren), Schwankschwindel (wie Bootsfahren) oder Benommenheitsschwindel), 2.) der *Dauer* des Schwindels (Schwindelattacken über Sekunden bis

Minuten wie bei der Vestibularisparoxysmie, oder Minuten bis Stunden wie bei Morbus Menière oder der vestibulären Migräne, in Abgrenzung zu Dauerschwindelsymptomen über mehrere Tage wie zum Beispiel bei der Neuritis vestibularis), sowie 3.) der Auslösbarkeit und Verstärkung des Schwindelsymptoms (zum Beispiel Auftreten in Ruhe, beim Gehen, bei Kopfdrehung oder Kopflagerung, oder in bestimmten Umgebungssituationen beim phobischen Schwankschwindel). Weiterhin sind typische Begleitsymptome ausschlaggebend für die Diagnosestellung – bei Morbus Menière beispielsweise Ohrdruck, Tinnitus, Hörveränderung, Geräuschempfindlichkeit, wackelnde Bilder, Übelkeit, Erbrechen und Fallen.

Nach Strupp *et al.* (2013) zählen zu den medikamentös behandelbaren Schwindelerkrankungen unter anderem der phobische Schwankschwindel, Neuritis vestibularis, Morbus Menière, Vestibularisparoxysmie, und zentral vestibuläre Formen wie zum Beispiel Episodische Ataxien, cerebelläre Stand- und Gangataxie, oder die vestibuläre Migräne. Ziel kausaler Therapieansätze bei vestibulären Schwindelsyndromen mit dem Leitsymptom episodisch auftretender Attacken ist die Reduktion der Attackenhäufigkeit, beziehungsweise längerfristig die Prävention von Attacken (vollständige Attackenfreiheit).

Trotz der hohen klinischen Relevanz ist die Versorgungssituation von Patienten mit Schwindelsyndromen noch immer unzureichend. Vor der Behandlung wird of keine exakte Diagnose gestellt, eine Vielzahl der Patienten erhält daher keine adäquate Therapie, und es kommt zu einer Fehl- und Überversorgung (Neuhauser, 2009; Rieger *et al.*, 2014). Zum Teil fehlen für bestehende Therapiekonzepte randomisierte kontrollierte klinische Studien, die die Wirksamkeit und Sicherheit der zu prüfenden Medikation belegen.

1.1.2 Grundprinzipien konfirmatorischer klinischer Studien

Prospektive randomisierte kontrollierte klinische Studien der Phase III haben zum Ziel, eine spezifische, klinisch relevante Fragestellung zu Therapieeffekten klar und eindeutig zu beantworten und verlässliche Evidenz bezüglich Wirksamkeit (Efficacy bzw. Effectiveness) und Sicherheit der zu prüfenden Intervention zu liefern. Die Integrität und Interpretation der Studienergebnisse hängt entscheidend davon ab, ob, abgesehen von einem adäquaten Studiendesign und einer sorgfältigen Studiendurchführung, das zugrunde liegende biometrische Analysekonzept bestimmten Qualitätsanforderungen genügt (ICH E9, 1998, Kap. 2). Im Gegensatz zu rein explorativen Studien wird bei konfirmatorischen Phase III Studien das Analysekonzept für primäre und ggfs. für vorab definierte sekundäre Schlüsselendpunkte *a priori* detailliert festgelegt. Hierfür beschreibt die Guideline ICH E9 (Statistical Principles for Clinical Trials, 1998) allgemeine statistische Prinzipien und Empfehlungen für die bei klinischen Studien relevanten statistischen Methoden. Weitere neuere internationale Leitlinien wie das im Jahre 2013 veröffentlichte SPIRIT-Statement* haben zum Ziel, Minimalstandards für Studienprotokolle zu etablieren, und fordern anhand von 31 Items bestimmte Inhalte, welche im Prüfplan berücksichtigt werden sollten

*SPIRIT: Standard Protocol Items: Recommendations for Interventional Trials

(Chan *et al.*, 2013a,b). Dazu gehört neben der präzisen Formulierung der primären und sekundären Studienziele, der Operationalisierung der primären und sekundären Outcomes, d. h. Variablen zur Messung der Wirksamkeit und Sicherheit der Therapie, sowie der statistischen Hypothese(n) die Spezifikation der zugehörigen statistischen Auswertungsstrategie insbesondere für die Hauptanalyse. Dies beinhaltet die Wahl geeigneter statistischer Testverfahren oder modellbasierter Analysen im frequentistischen oder Bayesianischen Kontext (vgl. SPIRIT Item 20).

Darüber hinaus werden im statistischen Abschnitt des Prüfplans weitere prospektiv geplante, konfirmatorische Analysen skizziert: Neben der präspezifizierten Hauptanalyse zum Nachweis der Überlegenheit einer experimentellen Therapie beinhaltet dies adjustierte Analysen für primäre (und sekundäre) Zielkriterien, bei denen wenige Baseline-Kovariablen ausgewählt werden, beispielsweise zur Berücksichtigung bekannter prognostischer Faktoren, nach denen randomisiert wurde (d. h. im Falle einer stratifizierten Randomisierung), oder vordefinierte Subgruppen-Analysen zur Untersuchung der Homogenität des geschätzten Behandlungseffekts. Weitere Details und Grundprinzipien zu adjustierten Analysen und Subgruppen-Analysen findet man unter anderem in den entsprechenden Guidelines CHMP (2015) und CHMP (2014a) der EMA[†].

1.1.3 Motivierendes Beispiel: die BEMED-Studie

Morbus Menière ist eine chronisch progrediente Erkrankung des Innenohrs, gekennzeichnet durch anfallartig auftretende Schwindelepisoden. Die Lebenszeitprävalenz liegt bei etwa 0.5% (Neuhauser, 2007). Die Erkrankung beginnt meist einseitig, im weiteren Krankheitsverlauf entwickelt sich meist eine bleibende Hörminderung auf dem betroffenen Ohr, und ca. 50% der Patienten entwickeln einen bilateralen Morbus Menière. Leitsymptome einer klassischen Menière-Attacke sind Minuten bis mehrere Stunden anhaltender akuter Drehschwindel mit einseitiger chronischer Hörminderung, Tinnitus oder Druckgefühl im betroffenen Ohr (Menière'sche Trias), sowie weitere typische Begleitsymptome wie zum Beispiel Übelkeit, Erbrechen, Geräuschempfindlichkeit, oder wackelnde Bilder vor den Augen. Ein Schwindelereignis kann von einem Drehschwindel in einen Schwankschwindel und später in eine abnehmende Gangunsicherheit oder Benommenheit übergehen, wobei die individuelle Wahrnehmbarkeit des Schwindelereignisses recht unterschiedlich ist. In manchen Fällen äußert sich die akute Menière-Attacke auch durch einen heftigen Schwankschwindel, oft in Zusammenhang mit Gangunsicherheit, selten auch verbunden mit plötzlichem Zu-Boden-Stürzen (drop-attack). Die Frequenz der Schwindelepisoden variiert stark über die Zeit, mit Phasen häufig auftretender Symptome und Beschwerdefreiheit beziehungsweise selten auftretender Episoden im Intervall. Weitere medizinische Details findet man unter anderem bei Brandt *et al.* (2004, Kap. 2.3) oder Strupp & Brandt (2008).

Eine rein symptomatische medikamentöse Therapie zur Minderung von Schwindelsymptomen wie Übelkeit und Erbrechen erfolgt mit Antivertiginosa. Eine prophylaktische

[†]European Medicines Agency

Langzeit-Therapie zielt auf die Reduzierung der Attackenfrequenz oder vollständige Beschwerdefreiheit. Mittel der Wahl ist aktuell das in Europa seit den 1970er Jahren zugelassene Betahistin-Dihydrochlorid, bei einer maximalen Tagesdosis von 48 mg. Dennoch fehlen bislang randomisierte placebo-kontrollierte klinische Studien nach heutigen Qualitätsstandards, die die Wirksamkeit von Betahistin belegen (Murdin *et al.*, 2016). Die Metaanalysen von Nauta (2014) sowie Della *et al.* (2006) geben Hinweise auf einen positiven Effekt von Betahistin bei Morbus Menière, berücksichtigen aber eine relativ geringe Anzahl an placebo-kontrollierten Studien, unter Verwendung eines ordinal skalierten Wirksamkeitsendpunkts für die in die Metaanalyse eingeschlossenen Studien (Whitehead & Jones, 1994). Ein Cochrane Systematic Review von 2001 weist auf die methodischen Schwächen bisheriger Therapiestudien mit Betahistin hin. Es gibt bislang keine klare Evidenz für eine positive therapeutische Wirkung von Betahistin bei Morbus Menière bzw. beim Menière'schen Symptomkomplex (James & Burton, 2001; James & Thorp, 2007; Harcourt *et al.*, 2014). Hinweise auf einen dosisabhängigen Effekt dieser medikamentösen Intervention sind bislang durch placebo-kontrollierte klinische Studien nicht ausreichend belegt (Lezius *et al.*, 2011).

1.1.3.1 Studienrationale und biometrisches Konzept

Aus diesem Grund wurde die BEMED-Studie (Medical treatment of *ME*nière's *D*isease with *BE*tahistine)[‡] initiiert, eine im akademischen Setting durchgeführte multizentrische, randomisierte, doppelblinde, placebo-kontrollierte, dreiarmige Phase III Therapieoptimierungsstudie im Parallelgruppendesign. Betahistin ist in Deutschland für die Behandlung des Morbus Menière in der Tagesdosierung bis 2×24 mg zugelassen (Standarddosis), nicht jedoch die in der BEMED-Studie untersuchte experimentelle Hochdosis von 3×48 mg pro Tag. Primäres Studienziel ist die Untersuchung des Effekts einer prophylaktischen längerfristigen Therapie mit Betahistin auf die Anzahl der akut auftretenden Menière-Attacken, genauer die Attacken*inzidenz*. Studienrationale ist die Hypothese eines dosisabhängigen Wirkmechanismus, das heißt es wird angenommen, dass eine hochdosierte Langzeit-Behandlung mit Betahistin über viele Monate einer niedrigeren Dosierung oder der Placebo-Intervention überlegen ist, gemessen an der Reduzierung der Attackenfrequenz über die Zeit. Zu den sekundären Wirksamkeitsendpunkten gehören die Dauer und Intensität der Attacken, diverse Parameter zur Messung der peripher vestibulären und audiologischen Funktion, sowie verschiedene krankheitsspezifische Lebensqualitäts-Scores.

Insgesamt wurden in 14 Studienzentren verteilt über ganz Deutschland 221 Patienten mit der Diagnose eines Morbus Menière randomisiert im Verhältnis 1:1:1 auf die drei Behandlungsgruppen

- Placebo,
- Standarddosis Betahistin (2×24 mg/Tag),

[‡]EudraCT Nr. 2005-000752-32, Protocol Code Nr. 04T-617, Principal Investigator und Sponsor am Klinikum der LMU München. Randomisierungsbeginn war 03/2008, Last-Patient Last-Visit war 11/2013.

- Hochdosis Betahistin (3×48 mg/Tag),

wobei die diagnostischen Kriterien entsprechend der Guideline der American Academy of Ophthalmology and Otolaryngology, Head and Neck Surgery (AAO-HNS, 1995) definiert wurden. Die individuelle Behandlungsdauer beträgt prüfplankonform 9 Monate. Im Anschluss an die Treatment-Phase erfolgt eine 3-monatige Follow-up Phase.

Hauptendpunkt ist die *Anzahl der Attacken* pro Zeiteinheit, definiert als 30-Tage Intervalle $t = 1, 2, \ldots, 9, \ldots, 13$ (gerechnet ab Therapiebeginn)[§]. Dieser muss bei der statistischen Analyse in Beziehung gesetzt werden zur *Anzahl an Beobachtungstagen* pro Zeiteinheit, interpretierbar als Dauer der Exposition beziehungsweise Anzahl der Tage mit dokumentierter (nicht fehlender) Attackeninformation. Die Anzahl an Beobachtungstagen in einem bestimmten Zeitintervall kann reduziert sein zum Beispiel aufgrund von vorzeitigem Studienabbruch (z. B. wegen Dropout oder Loss-to Follow-up), oder fehlender Dokumentation des Attackenstatus bei fortlaufender Beobachtungs- und Treatmentphase. Somit ist die patientenindividuelle Anzahl der Beobachtungstage (bewertbare Tage) pro Zeiteinheit definiert als die Länge des Zeitintervalls abzüglich der Anzahl der Tage mit fehlender Attackeninformation (*Missings*). Ist die Attacken-Dokumentation zwischenzeitlich fehlend und wird nach einer bestimmten Zeit wieder aufgenommen, so resultiert dies in *intermittierenden*, d. h. nicht-monotonen, Missings bezogen auf den primären Endpunkt. Handelt es sich um einen endgültigen Dokumentationsabbruch und somit um ein Fehlen des Attackenstatus ab einem bestimmten Zeitpunkt (verbunden mit oder ohne Studienabbruch), spricht man von *monotonen* Missings. Die letztgenannte Situation stellt ein klassisches 'Missing outcome data'-Problem im statistischen Sinne dar. Abbildung 1.1 zeigt exemplarisch den individuellen Attackenverlauf eines Patienten der BEMED-Studie mit regulärer Studiendauer von 12 Monaten. Im Intervall 7 und 8 ist die Anzahl an Beobachtungstagen kleiner als 30 aufgrund von intermittierenden Missings, da der Attackenstatus nicht kontinuierlich dokumentiert wurde. Ab Intervall 9 ist die Anzahl an Beobachtungstagen maximal (30 Tage), bevor diese im Intervall 13 reduziert ist wegen regulärem Beobachtungsende nach 362 Tagen.

Die Auswertung des primären Endpunkts beinhaltet die Messung des Unterschieds zwischen den drei Behandlungsgruppen[¶] bezüglich der Attackeninzidenz in den letzten 3 Monaten der 9-monatigen Behandlungsphase, und somit in den Zeitintervallen $t = 7, 8, 9$. Der prä-spezifizierte Bewertungszeitraum zur Untersuchung der Nullhypothese beinhaltet also einen Zeitraum von 90 Tagen und wird im folgenden als *Assessment-Periode* bezeichnet. Die der Hauptfragestellung zugrundeliegende globale Nullhypothese $H_{0,\,\text{global}}$ der BEMED-Studie lautet:

$H_{0,\,\text{global}}$: Es gibt keinen Unterschied in der Attackeninzidenz im (aggregierten) Zeitintervall 7, 8, und 9 zwischen den drei Behandlungsgruppen.

[§]Bei strenger äquidistanter Einteilung der Zeitachse würden sich bei einer prüfplankonformen Studiendauer von exakt 12 Monaten (365 Tage) somit 13 Intervalle ergeben.

[¶]Das Efficacy Outcome Measure ist definiert als die mittlere Attackeninzidenz in den letzten 3 Monaten der 9-monatigen Behandlungsphase.

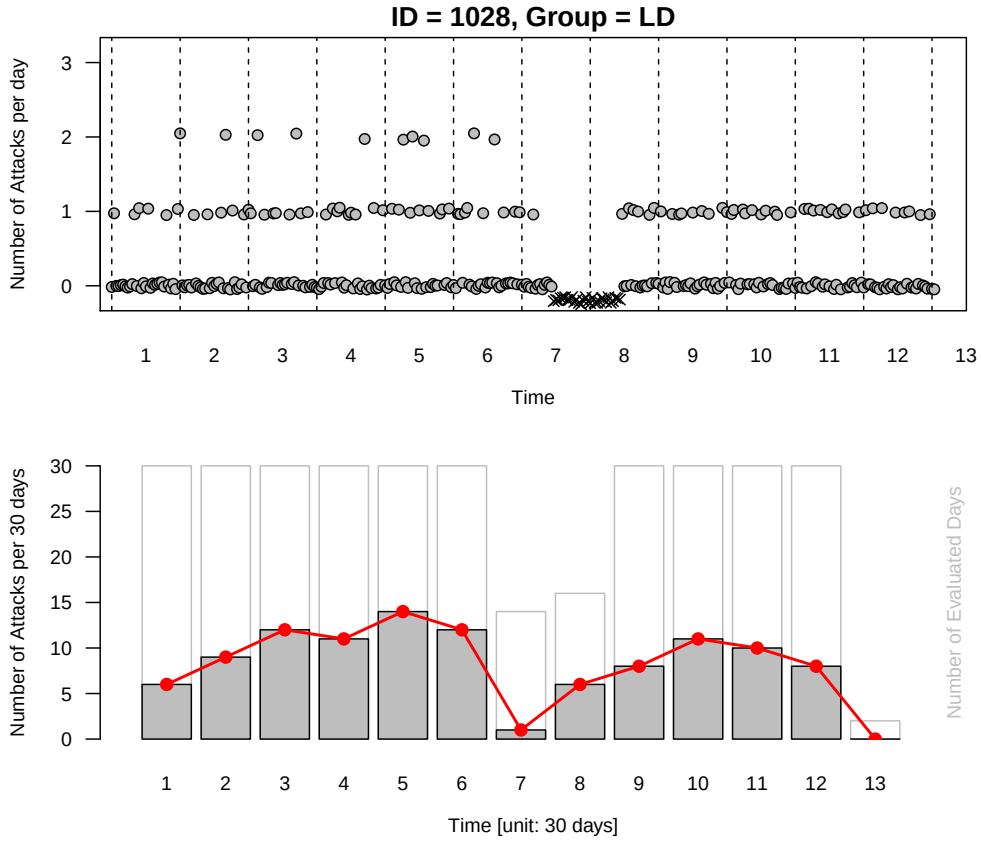

Abbildung 1.1: BEMED-Studie: Beispiel eines individuellen Attackenverlaufs (Treatment-Arm 'low-dose' (Standarddosis Betahistin)), bei unvollständiger Tagebuchdokumentation in der Assessment-Periode $\{7, 8, 9\}$. Oberer Plot: Rohdaten auf täglicher Basis, mit Tag 1 definiert als Beginn der Treatment-Phase (graue Punkte: Attackeninzidenz pro Tag; Kreuzchen: Tage ohne Dokumentation (zur besseren Darstellung wurde Jittering verwendet); senkrechte gestrichelte Linien kennzeichnen Zeitintervalle von 30 Tagen). Unterer Plot: Balkendiagramm mit individuellem Verlauf der Attackeninzidenz (rote Linie: Polygonzug) aggregiert pro Intervall. Die Anzahl bewertbarer Tage (Höhe der weißen Balken) mit Attackeninformation im Zeitintervall 7, 8 und 13 beträgt weniger als 30 Tage aufgrund von Missings.

Basierend auf dem Abschlusstestprinzip wird bei Einhaltung des Signifikanzniveaus von $\alpha = 5\%$ auf das Vorhandensein eines globalen Treatment-Effekts getestet (globale Nullhypothese $H_{0,\,\text{global}} = H_{01} \cap H_{02} \cap H_{03}$), gefolgt von Paarvergleichen für die drei Nullhypothesen H_{01}: Hochdosis vs. Standarddosis, H_{02}: Hochdosis vs. Placebo, H_{03}: Standarddosis vs. Placebo. Kann die globale Nullhypothese $H_{0,\,\text{global}}$ nicht abgelehnt werden, können keine validen Paarvergleiche unter Einhaltung des 5%-Niveaus durchgeführt werden.

Die methodischen Details zur Auswertungsstrategie der BEMED-Studie, insbesondere zur modellbasierten Primäranalyse, findet man im Statistischen Analyseplan, siehe Anhang ab Seite 44.

1.1.3.2 Anwendungsbeobachtung "Betahistin"

Grundlage für Studiendesign und -planung der BEMED-Studie waren die Verlaufsdaten von 112 Menière-Patienten einer von Strupp *et al.* (2008a) am Klinikum der LMU München im Vorfeld durchgeführten Anwendungsbeobachtung, eine monozentrische, unverblindete, nicht-kontrollierte, nicht-randomisierte 2-armige Pilotstudie, bei der die Überlegenheit der in der BEMED-Studie untersuchten höheren Dosierung von Betahistin (3×48 mg pro Tag, $N = 62$ Patienten) im Vergleich zu einer niedrigeren Dosierung ($N = 50$ Patienten; 3×16 mg pro Tag, $N = 21$; bzw. 3×24 mg pro Tag, $N = 29$) bei einer Behandlungsdauer von 12 Monaten nachgewiesen werden konnte. Die Dosierung in jeder Gruppe erfolgte konstant über die gesamte 12-monatige Behandlungsdauer. Hauptendpunkt der Studie war die Anzahl der Menière-Attacken pro Monat (naiv gemittelt jeweils über ein 3-Monats-Intervall), gemessen wurde retrospektiv 3 Monate vor Therapiebeginn $t = 0$, und zu den Zeitpunkten $t = 3, 6, 9$ und 12 Monate.

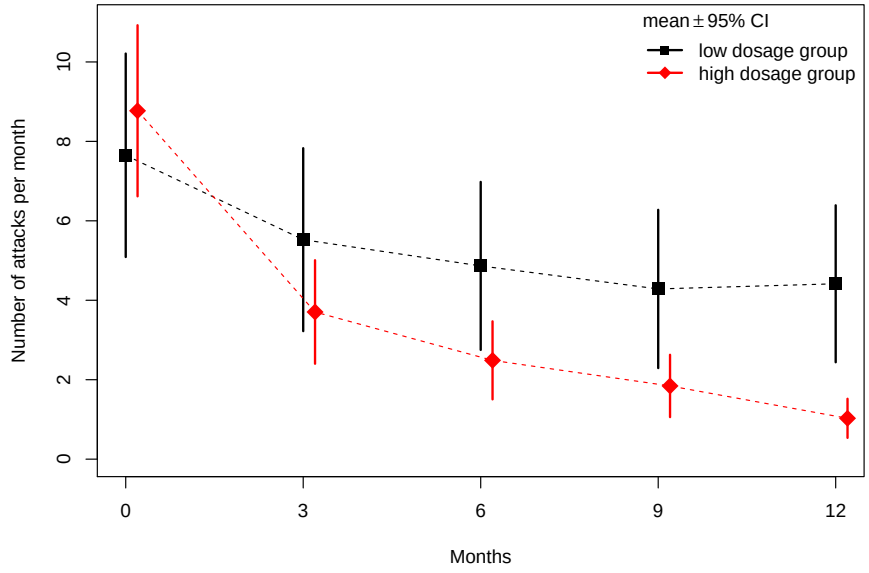

Abbildung 1.2: Anwendungsbeobachtung aus Strupp *et al.* (2008a). Einfache deskriptive Darstellung des Effekts von Betahistin-Dihydrochloid auf die Attackeninzidenz bei insgesamt 112 Patienten mit Morbus Menière. 2 Behandlungsgruppen (Comparator vs. experimentelle Dosis): zugelassene Standarddosis 16 bzw. 24 mg dreimal täglich (schwarz), vs. Hochdosis 48 mg dreimal täglich (rot). Dargestellt ist die mittlere Anzahl der Attacken (Punktschätzer ± 95% KI (punktweise)) der letzten 3 Monate vor Studienbeginn (Baseline-Zeitpunkt 0), sowie zu den Zeitpunkten 3, 6, 9 und 12 Monate nach Therapiebeginn.

In Abbildung 1.2 ist deskriptiv für beide Behandlungsgruppen die zu den fünf Messzeitpunkten berechnete mittlere Attackenfrequenz dargestellt, zusammen mit punktweise 95%-Konfidenzintervallen. Mit einer modellbasierten Analyse (ohne Berücksichtigung der Anzahl bewertbarer Tage pro Erhebungszeitraum) konnte die signifikante Abnahme der

Attackeninzidenz über die Zeit nachgewiesen werden. Gleichzeitig zeigte sich im Zeitverlauf eine signifikant höhere Reduktion der Attackeninzidenz in der Hochdosis-Gruppe im Vergleich zur niedrigeren Dosierung (bei einer Treatment-Dauer von 12 Monaten). Für Details zu den Ergebnissen der Studie mit explorativem Charakter sei auf den Artikel von Strupp *et al.* (2008a) verwiesen.

Eine methodische Limitation dieser Studie ist das Ignorieren der Attacken-Daten von 16 Dropouts (Treatment-Dropouts bzw. Analyse-Dropouts (Studienabbrecher)), also Patienten ohne vollständig beobachtete Attackeninformation. Für die Hauptanalyse wurden 112 Patienten mit kompletten Attacken-Verläufen bis Monat 12 berücksichtigt ('Completers'). Diese *Complete-Case-Analyse* liefert ein valides Schätzergebnis unter der Annahme *Missing Completely At Random* (MCAR) bezogen auf den primären Endpunkt. Gilt MCAR, so ist die Wahrscheinlichkeit für Missing (d. h. das Fehlen einer einzelnen Beobachtung) weder von den beobachteten noch von den unbeobachteten Werten abhängig. Einen guten Überblick zu den verschiedenen Missingness-Mechanismen liefern z. B. Carpenter & Kenward (2007) oder Little & Rubin (2002). Eine weitere Limitation ist unmittelbar zurückzuführen auf das Studiendesign einer Anwendungsbeobachtung: Die Datenerhebung erfolgte unter nicht-kontrollierten Bedingungen. Zusätzlich zu der fehlenden Verblindung und des damit verbundenen Verzerrungspotentials sowie der fehlenden Randomisierung zu den beiden alternativen Behandlungen, ist durch die nicht vorhandene Placebo-Gruppe eine direkte oder indirekte Abschätzung des Placebo-Response nicht möglich. Die Bedeutung einer Placebo-Intervention insbesondere bei chronischen symptomatischen Erkrankungen wie Schwindel, dessen natürlicher Krankheitsverlauf (ohne Intervention) gekennzeichnet ist durch spontan rezidivierende Symptome, aber auch durch vollständige Attackenfreiheit über längere Zeitphasen, wird unter anderem von Hamill (2006) am Beispiel von Morbus Menière diskutiert. Für weitere medizinische Details zur Natural History bei Morbus Menière sei auf Perez-Garrigues *et al.* (2008) verwiesen.

Die in dieser offenen Vorstudie generierte Hypothese einer dosisabhängigen Reduzierung der Attackenrate über die Zeit war der Anlass für die Initiierung der konfirmatorischen, verblindeten, randomisierten, 3-armigen placebo-kontrollierten BEMED-Studie, die den Wirksamkeitsvergleich zweier Dosierungen von Betahistin mit Placebo ermöglicht.

1.2 Zielsetzung dieser Arbeit

Ziel dieser Dissertation ist es, biometrische Prinzipien bei der Durchführung von Therapiestudien bei Schwindelerkrankungen mit dem Leitsymptom rezidivierender Attacken vorzustellen, welche für den statistischen Teil des Prüfplans im Rahmen der Studienplanung, aber insbesondere für die Entwicklung und Ausarbeitung des Statistischen Analyseplans (SAP) relevant sind, und somit 1.) nicht datengeleitet und 2.) ohne Kenntnis der Gruppenzugehörigkeit. Im Fokus stehen prospektive, *verblindete*, individuell randomisierte kontrollierte klinische Studien (Randomized Controlled Trials, RCTs) im Parallelgruppen- oder Crossover-Design, die die Wirksamkeit – Efficacy beziehungsweise Effectiveness – einer

Therapie anhand von Zähldaten über die gesamte individuelle Studiendauer hinweg messen. Zudem gehen wir in dieser Arbeit von *konfirmatorischen* Studien aus, deren Ziel es ist, die *Über*legenheit (Superiority) einer experimentellen Therapie im Vergleich zu einer Standardtherapie (aktive Kontrolle) und/oder Placebo nachzuweisen. Es soll also der Wirksamkeitsnachweis erbracht werden, dass die experimentelle Therapie der Vergleichstherapie überlegen ist hinsichtlich des vordefinierten Hauptendpunkts, d. h. zum Beispiel zu einer signifikanten und klinisch relevanten Reduktion der Attackeninzidenz innerhalb einer präspezifizierten Assessment-Periode führt.

Die Erhebung der Rohdaten für den primären Endpunkt erfolgt bei diesem Krankheitsbild häufig durch Patiententagebücher. Der adäquate Umgang mit diesem sogenannten Patient-Reported Outcome (PRO)-Instrument bringt weitere methodische Herausforderungen mit sich (FDA, 2009; CHMP, 2005; Calvert *et al.*, 2013; Izem *et al.*, 2014).

In dieser Arbeit werden die wichtigsten Grundprinzipien zur statistischen Analyse von longitudinalen Zähldaten bei RCTs vorgestellt, und am Beispiel der BEMED-Studie die Problemstellung aus der Sicht des verantwortlichen Biometrikers erläutert sowie Lösungsansätze aufgezeigt.

2 Methodik für Schwindelstudien mit longitudinalen Zähldaten

2.1 Studiensetting und Designaspekte

Prospektive klinische Studien zu Schwindelerkrankungen mit dem Leitsymptom episodisch auftretender Attacken haben häufig das primäre Ziel, die Wirksamkeit einer prophylaktischen Therapie in Bezug auf die Attackenfrequenz zu untersuchen. Primäres Outcome ist hierbei die Anzahl der Attacken beziehungsweise die Attackeninzidenz pro definierter Zeiteinheit. Zur Quantifizierung des Therapieeffekts werden Zähldaten über die gesamte individuelle Behandlungsperiode und in der Regel darüber hinaus über einen bestimmten Follow-up Zeitraum kontinuierlich erhoben. Konfirmatorische Therapiestudien (Randomized Controlled Trials) zu chronischen Schwindelerkrankungen wie Morbus Menière oder Vestibuläre Migräne werden üblicherweise im Parallelgruppen-Design durchgeführt, wenn der Effekt einer experimentellen symptomatischen Langzeit-Therapie über viele Monate (in der Regel 6 Monate und länger) im Vergleich zu einer Standardtherapie und/oder Placebo untersucht werden soll, und der natürliche Verlauf der Attackeninzidenz über größere Zeiträume starken Schwankungen inklusive Phasen vollständiger Beschwerdefreiheit unterliegt. Wie bei der BEMED-Studie kann es sich bei der Standardtherapie auch um die dem experimentellen Arm entsprechende zugelassene Standarddosierung handeln, deren Wirksamkeit mit der einer höheren Dosierung oder Placebo verglichen werden soll. Angenommen wird hierbei ein sich über die Zeit eher langsam einstellender Behandlungseffekt, im Gegensatz zu einem relativ kurz nach Therapiebeginn eintretender starker Effekt, welcher im weiteren Verlauf über die gesamte Studiendauer eher stabil bleibt.

Bei chronischen Schwindelerkrankungen mit geringer Prävalenz, wie zum Beispiel Vestibularisparoxysmie oder EA2, kommt das Crossover-Design in Betracht, sofern für die geplante Behandlungsdauer von keiner nennenswerten Progression der Erkrankung und somit von einer Stabilität der Symptome ohne Intervention ausgegangen werden kann. Zudem eignet sich dieses Studiendesign bei Studien mit kürzerer Behandlungsdauer (z. B. Wochen bis wenige Monate) unter der Annahme eines sich eher schnell einstellenden Behandlungseffekts, welcher nach Absetzen der Therapie das Zielkriterium (Häufigkeit der Symptome) idealerweise wieder auf das Baseline-Niveau zurückkehren lässt. Im klassischen 2-Perioden Crossover-Design zur Beurteilung der Wirksamkeit zweier Behandlungen A und B werden Patienten zu den zwei Behandlungssequenzen A/B und B/A randomisiert. Zwischen den

beiden Behandlungsperioden erfolgt eine ausreichend lange Wash-out Phase zur Minimierung des Carryover-Effekts. Da jeder Patient beide Behandlungen A und B erhält, somit als seine eigene Kontrolle dient, besitzt das Crossover-Design eine höhere Effizienz im Vergleich zu einer entsprechenden Studie im Parallelgruppen-Design, welche pro Patient Messdaten entweder nur unter der Behandlung A oder B liefert. Geht man davon aus, dass die Messfehler-Varianz (within-patient error) innerhalb eines Patienten in der Regel deutlich geringer ist im Vergleich zu der im Parallelgruppen-Design relevanten interindividuellen Variabilität (between-subject error), so kann der Unterschied in der Fallzahl zwischen Crossover- und Parallelgruppen-Design bei sonst gleichen Planungsannahmen erheblich sein. Im Parallelgruppen-Design werden demnach mehr Patienten benötigt, um dieselbe Power zu erreichen wie im Crossover-Design.
Für eine ausführliche Einführung in die Methodik des Crossover-Designs sei auf Senn (2002) oder Jones & Kenward (2014) verwiesen. Ein Beispiel für eine placebo-kontrollierte 2-Perioden 2-Treatment Crossover-Studie, in der die Wirksamkeit von 4-Aminopyridin bezüglich der Attackenfrequenz bei EA2 untersucht wird, findet man im medizinischen Artikel Strupp *et al.* (2011).

Wir präsentieren in den folgenden Abschnitten die zugrundeliegende Methodik für die Auswertung von longitudinalen Zähldaten (tagebuchbasierten Attackendaten) anhand von konfirmatorischen RCTs zu Schwindelerkrankungen im Parallelgruppen-Design. Viele der getroffenen Aussagen und Problemstellungen hinsichtlich der Analyse von Attackendaten gelten in angepasster Form für RCTs im Crossover-Design.

Die in dieser Arbeit vorgestellte Methodik zur Modellierung von Zähldaten ist aber auch für Beobachtungsstudien aus der Routineversorgung relevant, welche retrospektiv Verlaufsdaten von Patienten mit oder ohne Therapie zu individuellen Erhebungszeitpunkten, definiert in der Regel durch Visiten, und individuell variierender Behandlungs- sowie Beobachtungsdauer erheben. Exemplarisch sei hier die Beobachtungsstudie von Hüfner *et al.* (2008) genannt, die für Patienten mit Vestibularisparoxysmie den Unterschied in der Attackeninzidenz mit und ohne Medikation untersucht.

2.2 Patient-Reported Outcome (PRO) als Efficacy-Endpunkt

In diesem Abschnitt werden die wichtigsten methodischen Eigenschaften bei der Messung von Krankheitssymptomen bei Schwindelstudien mit longitudinalen Zähldaten vorgestellt. Es werden die Herausforderungen bei der Erhebung der Attackendaten anhand von papierbasierten Patiententagebüchern erläutert, und die Problematik bei der Ableitung der für die Hauptanalyse relevanten Efficacy-Daten anhand eines Beispiels kurz skizziert.

2.2.1 Hintergrund

Die Definition eines geeigneten und klinisch relevanten primären Endpunkts (Outcome Measure), welcher die einer klinischen Prüfung zugrundeliegende Hauptfragestellung zur Wirksamkeit abbildet, und mit dem der Effekt der Intervention valide und reliabel gemessen werden kann, erfolgt im Rahmen der Studienplanung. Neben den klassischen Zielgrößen, die auf einer subjektiven Einschätzung oder Interpretation des Beobachters (des Prüfarztes) beruhen, sog. *Clinician* bzw. *Observer Reported Outcomes*, oder objektiv messbare Kriterien (Performance Outcomes), die bei bestimmten Untersuchungen erfasst werden (z. B. Messung der Ganggeschwindigkeit, Laborparameter, biologische, physiologische oder apparativ gemessene Parameter), gewinnen in der patientenorientierten klinischen Forschung patientenberichtete Endpunkte, *Patient Reported Outcomes (PROs)*, zunehmend an Bedeutung. PROs bezeichnen unterschiedliche Konzepte zur Erhebung von *subjektivem* Krankheitsempfinden (z. B. Symptom-Status, Aspekte der Funktionsfähigkeit, die mit dem Krankheitszustand in Zusammenhang stehen, oder subjektive Wahrnehmung von bestimmten Veränderungen des eigenen Gesundheitszustandes im Zeitverlauf aufgrund einer Therapie), Patientenzufriedenheit oder krankheitsspezifischen Quality-of-Life Zuständen (CHMP, 2005, 2014b), welche vom Patienten individuell wahrgenommen und dokumentiert werden. PROs messen somit zentrale Aspekte der Krankheitslast aus Patientensicht. Hierfür bedarf es geeigneter Instrumente, die für den Einsatz in klinischen Studien, für die entsprechende Patientenpopulation, und für eine spezifische Studienfragestellung geeignet sind. Die Guideline zu PROs der FDA (2009) gibt eine umfassende Übersicht über die nötigen Anforderungen im Rahmen von Zulassungsstudien.* Beispiele für PRO Instrumente, mit denen die Wirksamkeit einer bestimmten Therapie gemessen wird, sind Selbstbeurteilungsfragebögen, die der Patient ohne Unterstützung durch einen Interviewer alleine ausfüllt, oder bestimmte Selbsteinschätzungsskalen. Bei verschiedenen Schwindelerkrankungen werden häufig symptom- oder krankheitsspezifische Patientenfragebögen zur Erhebung vestibulärer Scores verwendet. Beispiele hierfür sind der Dizziness Handicap Inventory (DHI), die Vestibular Disorders Activities of Daily Living (VDADL) Skala, Tinnitus-Fragebögen, oder der Menière's Disease Patient Oriented Severity Index (Gates & Verrall, 2005).

Die Validität dieser PRO Instrumente in einem bestimmten Studiensetting, d. h. für eine bestimmte Indikation sowie Studienpopulation, zur Messung des Effekts einer Intervention ist nicht immer gegeben, vgl. hierzu z. B. der systematische Review von Fong *et al.* (2015) zu gängigen PRO Instrumenten bei vestibulären Erkrankungen. Wird in einer konfirmatorischen RCT die Wirksamkeit einer Therapie anhand eines PRO Instruments gemessen, muss dieses bestimmten Qualitätsanforderungen genügen. Zu den Gütekriterien gehören Reliabilität, Validität, die Fähigkeit, Veränderungen der Krankheitssymptome im Zeitverlauf abbilden zu können (Responsiveness), sowie Interpretierbarkeit der abgeleiteten Endpunkte entsprechend der Studienziele (FDA, 2009). Green *et al.* (2007) diskutieren

*Die FDA lässt für den Zulassungsprozess objektive Parameter und PROs als primären Efficacy-Endpunkt zu, nicht aber Endpunkte zur gesundheitsbezogenen Lebensqualität (Health-related Quality-of-life, HRQoL), welche nur begleitend als sekundäre Zielkriterien erhoben werden.

am Beispiel des Morbus Menière die Herausforderungen eines geeigneten krankheitsspezifischen QoL Instruments, welches die für Schwindelerkrankungen wie Morbus Menière typischen Fluktuationen der Symptome, insbesondere der Attackeninzidenz, über die Zeit valide abbildet. Eine gute Übersicht über Studien mit über die gesamte Studiendauer kontinuierlich gemessenen PRO Daten und zugehörige methodische Besonderheiten liefert der Review-Artikel von Fairclough (2004), sowie Bell & Fairclough (2014), Cappelleri & Bushmakin (2014) oder Kammerman & Grosser (2014). Im Vergleich zu Studien mit objektiv messbaren Endpunkten haben Studien mit einem patientenberichteten primären Endpunkt ein höheres Verzerrungspotential bei fehlender Verblindung (Wood *et al.*, 2008). Der Cochrane Review von Hróbjartsson & Gøtzsche (2010) liefert zudem Hinweise auf einen höheren Placebo-Response für RCTs mit Placebo Intervention, falls die Wirksamkeit der Behandlung anhand von PRO Daten gemessen wird.

2.2.2 Tagebuch-basierte Attackendaten

2.2.2.1 Schwindeltagebuch als PRO Instrument

Bei Therapiestudien zu chronischen Schwindelerkrankungen mit anfallartigen akuten Ereignissen wie z. B. bei Morbus Menière, EA2, Vestibulärer Migräne oder Vestibularisparoxysmie ist ein papierbasiertes Patiententagebuch bzw. Attackenkalender ein wichtiges Instrument zur Erhebung von PRO Daten. Es ermöglicht das Monitorieren und die selbständige, kontinuierliche Dokumentation der episodisch auftretenden, subjektiv wahrgenommenen Schwindelsymptome (Attacken) über die gesamte individuelle Studiendauer. Durch diese vom Patienten in seiner häuslichen Umgebung auf *täglicher* Basis durchgeführte Form der Datenerhebung können auftretende Schwindelepisoden aus Patientensicht über längere Zeiträume erfasst, und somit Krankheitsverläufe, Phasen kompletter Remission (Attackenfreiheit) sowie eventuelle Krankheitszyklen abgebildet werden. Die tägliche Beurteilung des Symptomstatus ('Symptom-Tracking') und ggfs. die Beschreibung aufgetretener Ereignisse und deren charakteristischer Eigenschaften (Kovariablen einer Attacke) minimiert den Recall-Bias, erfordert allerdings ein hohes Maß an Compliance (Hamill, 2006). Der Einsatz von papierbasierten Patiententagebüchern als Instrument zur Erhebung der Wirksamkeitsdaten bei einer lang andauernden Therapie ist insbesondere bei multizentrischen, wissenschaftsinitiierten Studien ohne Alternative, wird aber durchaus kritisch diskutiert (vgl. z. B. Stone *et al.*, 2002). Nach Stone *et al.* (2003) sind die Compliance und auch die Validität der resultierenden PRO Daten bei elektronischen Tagebüchern deutlich höher im Vergleich zu papierbasierten Tagebüchern. Eine gute Übersicht zu Grundprinzipien bei der Entwicklung eines Tagebuch-Instruments und zu dessen Validitätsnachweis liefert z. B. Gater *et al.* (2015).

Konfirmatorische Studien der Phase III setzen für die jeweilige Erkrankung und Patientenpopulation reliable und validierte PRO Instrumente zur Erhebung der Symptome voraus. Die Entwicklung eines solchen Tagebuch-Instruments zur Messung des Behandlungseffekts, welcher auch die Dokumentations-Last aus Patientensicht berücksichtigt, ist ein

längerer Prozess und birgt methodische Herausforderungen, ist aber von entscheidender Bedeutung. Das im klinischen Alltag aufgrund klinischer Erfahrungen bzw. Expertenwissen entwickelte und routinemäßig eingesetzte Patiententagebuch für eine eher pragmatische Dokumentation des individuellen Behandlungsverlaufs ist nicht zwangsläufig geeignet zur Ableitung von interpretierbaren und patientenrelevanten Efficacy-Daten im kontrollierten Studiensetting. Die psychometrischen Eigenschaften der bei RCTs zu Schwindelsyndromen eingesetzten Attackenkalender mit meist ereignisorientierter Dokumentation und einer Recall-Periode von maximal einem Tag sind bislang nicht ausreichend wissenschaftlich untersucht.

2.2.2.2 Tagebuch-Assessment: Attackendefinition und Endpunkt-Ableitung

Analog zu Migränestudien (CHMP, 2007) wird in RCTs zu Attackenschwindel als patientenrelevanter primärer Endpunkt in der Regel die Anzahl der Attacken pro definierter Zeiteinheit, als sekundäre Efficacy-Endpunkte die Attackenstärke (kategorial) sowie die Dauer verwendet (vgl. Abschnitt 1.1.3.1 für die BEMED-Studie). Die Anzahl der Attacken pro Zeiteinheit liefert einen Hinweis auf die individuelle Symptomschwere, berücksichtigt allerdings nicht die Tatsache, dass eine geringe Anzahl von starken und/oder lang anhaltenden Schwindelepisoden im Vergleich zu einer hohen Anzahl an milden Attacken (bezogen auf Stärke oder Dauer) subjektiv unterschiedlich bewertet werden kann (James & Burton, 2001; AAO-HNS, 1995). Bei Schwindelerkrankungen wie z. B. Morbus Menière oder Vestibulärer Migräne wird auf dem Patiententagebuch zusätzlich die Attackenart (Dreh- oder Schwankschwindel, Gangunsicherheit, oder Benommenheit) abgefragt. Für Morbus Menière sind weiterhin die während einer Attacke auftretenden typischen Begleitsymptome wie Tinnitus, Ohrdruck, Änderungen des Hörvermögens, sowie z. B. Geräuschempfindlichkeit, wackelnde Bilder, Übelkeit, Erbrechen oder Fallen klinisch relevant. Die Dokumentation komplexer Krankheitssymptome setzt ein hohes Maß an Compliance und Verstehen der abgefragten Items und Antwort-Optionen, sowie ein regelmäßiges Review durch den Prüfarzt im Studienverlauf voraus. Insbesondere steigt für den Patienten der Dokumentationsaufwand in Abhängigkeit vom individuellen Symptomstatus. Bei Schwindelsyndromen mit sehr kurzen, Sekunden bis wenige Minuten anhaltenden Episoden wie bei Vestibularisparoxysmie (Hüfner *et al.*, 2008) und generell bei mehreren Ereignissen pro Tag ist in der Praxis eine quantitative Dokumentation des Attackenstatus pro Tag erschwert. Dies kann zu unerwünschten, nicht prüfplankonformen qualitativen Angaben oder unpräzisen Schätzungen der Anzahl der Attacken auf dem Tagebuch und somit zu einer deutlichen Reduktion der Qualität der PRO Daten führen.

Das *Attacken-Counting*, d. h. die valide Ableitung der für die Primäranalyse entscheidenden Zähldaten anhand der auf dem Tagebuch dokumentierten *Rohdaten* ist nicht trivial. In Abhängigkeit von der Komplexität der zugrundeliegenden Erkrankung und des Designs des krankheitsspezifischen Tagebuchs ist es häufig notwendig, die Original-Eintragungen verblindet zu verifizieren, indem die Patienten-Ratings einer objektiven, kriterienorientierten Bewertung durch ein Endpoint Assessment Committee unterzogen werden. Der Prozess der Tagebuch-Evaluation und die Operationalisierung einer Attacke ist vor der

Entblindung festzulegen und erfolgt in der Regel anhand einer Standard Operating Procedure oder eines Consensus Dokuments[†]. Zu definieren sind unter anderem Regeln bezüglich der Abgrenzbarkeit von mehreren aufeinanderfolgenden Attacken, beispielsweise durch die a priori Festlegung einer attackenfreien Phase von z. B. 48 Stunden zwischen zwei aufeinanderfolgenden Attacken, um eine tagübergreifende Attacke langer Dauer oder Episoden mit ggfs. kurzen Unterbrechungen der Symptome von zwei zu wertenden Attacken unterscheiden zu können (vgl. EMA Guideline (2007) für RCTs zu Migräneattacken). Gegebenenfalls müssen Kriterien festgelegt werden hinsichtlich des Umgangs mit vorausgehenden bzw. abklingenden Symptomen (z. B. Aura-Phänomene), die der eigentlichen Attacke zuzuordnen sind und somit kein neues, zu wertendes Ereignis im Sinne der primären Fragestellung darstellen, sowie der Umgang mit sehr kurzen wiederkehrenden Episoden über ein bestimmtes Zeitintervall (Rezidiv). Im Gegensatz zu RCTs bei Kopfschmerz-Migräne, für die bereits entsprechende Consensus-Dokumente und Guidelines mit Definitionen für Efficacy Outcomes existieren (CHMP, 2007; IHS *et al.*, 2012; Silberstein *et al.*, 2008), gibt es für RCTs zu Schwindelsyndromen bislang keine vergleichbaren Guidelines, die sich mit der Operationalisierung und Ableitung der über Patiententagebücher dokumentierten Attackenfrequenz als primäres Wirksamkeitskriterium beschäftigen. Für Morbus Menière gibt die AAO-HNS (1995) Guideline sowie der Cochrane Systematic Review (James & Burton, 2001) lediglich Hinweise zur qualitativen Beurteilung des Symptomstatus. Gates (2000) diskutiert das Problem der Erhebung des Attackenstatus bei Morbus Menière und alternative Strategien der Definition des primären Outcomes, wie z. B. die Messung der symptomfreien Tage pro Zeiteinheit anstatt der Wertung von Attacken und zugehörigen Schwerestufen.

Grundsätzlich stellt bei Morbus Menière die Differenzierung zwischen organisch bedingten Symptomen einer Menière-Attacke und im Sinne der Studienziele nicht zu wertenden, vom Patienten dokumentierten Begleitsymptomen oder Dauersymptomen (z. B. mehrtägiges Benommenheitsgefühl oder Gangunsicherheit geringer Stärke) eine besondere Herausforderung dar. Abbildung 2.1 zeigt das in der BEMED-Studie verwendete Patiententagebuch am Beispiel einer einzelnen ausgefüllten Kalenderseite mit abgrenzbaren Schwindelattacken, und skizziert den komplexen Prozess der verblindeten Ableitung der für die primären und sekundären Studienziele benötigten Attackendaten anhand der dokumentierten Rohdaten. Grundlage einer möglichst validen und reliablen, manuellen Bewertung der dokumentierten Schwindelsymptome auf täglicher Basis war die im Rahmen der BEMED-Studie entwickelte SOP (Fischer *et al.*, 2014), welche pro Tag multiple Items für die Ableitung von Attacken verwendet, und unter anderem eine eine hierarchische Ordnung[‡] für verschiedene Schwindeltypen vorsieht. Hierbei wird auf Patienten-Ebene jeder einzelne Tag unter Beobachtung klassifiziert als Tag mit Null, einer oder mehreren Attacken, oder als Tag mit nicht bewertbarem Attackenstatus (Missing). Für weitere Details hinsichtlich der Entscheidungsregeln und des zugrundeliegenden Konzepts für den

[†]Ziel wäre beispielsweise ein rudimentäres algorithmisches Regelwerk, so dass anhand der Original PRO Daten durch Programmierung die für die Analyse relevanten Schwindelereignisse abgeleitet werden können.

[‡]Hierarchische Ordnung (Schwerestufe von hoch nach niedrig): Drehschwindel, Schwankschwindel, Gangunsicherheit, Benommenheit

Hauptendpunkt sei auf die SOP '*Diary Assessment*' von Fischer *et al.* (2014) verwiesen, welche ein offizieller Bestandteil des SAPs der BEMED Studie darstellt.

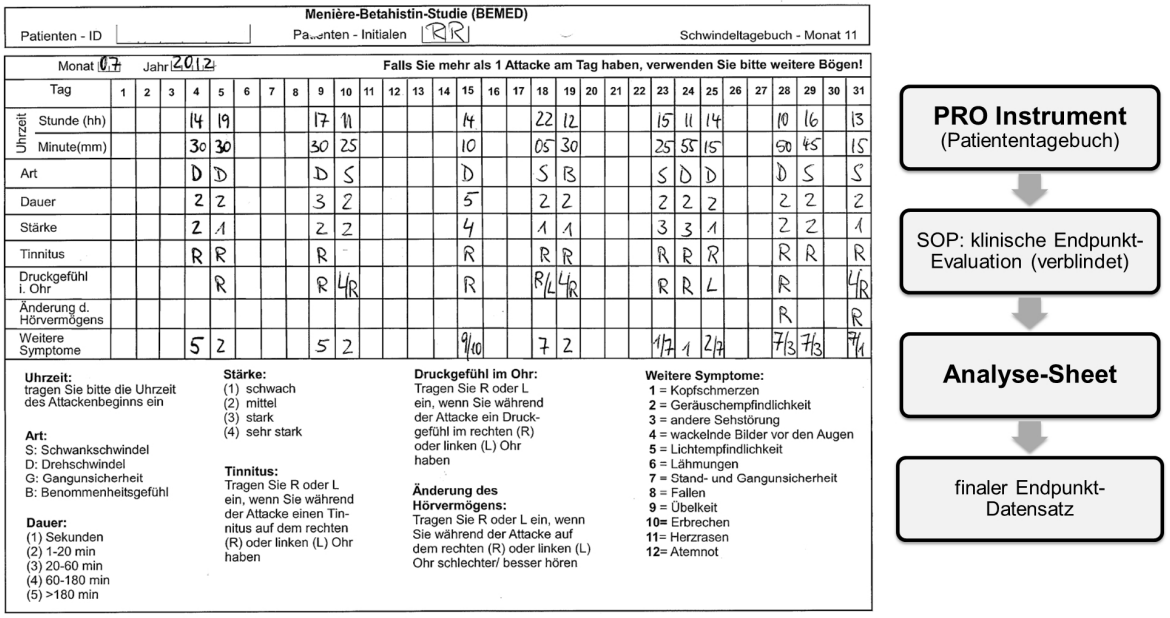

Abbildung 2.1: Tagebuch-Assessment durch ein zentrales Endpoint Adjudication Committee: Prozess der manuellen Attackenbewertung gemäß einer Standard Operating Procedure (SOP) am Beispiel der BEMED-Studie: verblindete Evaluation der anhand eines Patienten-Tagebuchs (Attackenkalender) auf täglicher Basis dokumentierten Schwindelsymptome (Rohdaten) zur standardisierten Ableitung des primären Zielkriteriums (Anzahl der Menière-Attacken pro vordefinierter Zeiteinheit) und Generierung des Efficacy-Datensatzes für die statistischen Analysen.

Die methodischen Herausforderungen bei der Ableitung von komplexen Wirksamkeitsendpunkten anhand von papierbasierten Patiententagebüchern mit *täglicher* oder *Ereignis-orientierter* Dokumentation lassen sich am Beispiel von Morbus Menière wie folgt zusammenfassen:

- Umgang mit ggfs. mehrtägigen Ereignisclustern bzw. "Musterbildern", sowie Klassifikation von Begleitsymptomen im zeitlichen Umfeld einer einzelnen Attacke; Differenzierung zwischen Rezidiv (zeitliche Unterbrechung) und neu aufgetretenem Ereignis im Sinne des interessierenden Zielkriteriums
- Umgang mit dokumentierten Dauersymptomen
- Umgang mit fehlenden hinreichenden Items (Kovariablen einer Attacke, z. B. zeitliche Charakteristik, Schwindelart, -stärke)
- Generierung von monotonen oder intermittierenden Missings: Umgang mit Dokumentationslücken, d. h. Zeiträume ohne Tagebuch-Dokumentation, oder zwischenzeitlich nicht bewertbaren Zeiträumen aufgrund unzureichender Dokumentationsqualität bzw. nicht eindeutig interpretierbarer Rohdaten

- Differenzierung zwischen symptomfreien (attackenfreien) Zeiträumen und Zeiträumen mit fehlender Dokumentation (z. B. fehlende Kalenderseiten) aufgrund mangelnder Compliance bezüglich des Tagebuch-Instruments
- Festlegung eines präspezifizierten Zeitintervalls für die *Aggregation* der bewerteten Attacken, und somit die Definition und Ableitung des Efficacy Endpunkts (Anzahl der Ereignisse pro Zeiteinheit).

2.3 Datenstruktur und Mixed-Effects Modellierung

2.3.1 Aggregation von Zähldaten

Die Aggregation der nach SOP abgeleiteten, auf täglicher Basis dokumentierten Attacken-Ereignisse und Beobachtungstage (d. h. bewertbare Tage mit bekanntem Attackenstatus, Tage unter Risiko) erfolgt mittels Summation über präspezifizierte Zeitintervalle zur Generierung von longitudinalen Zähldaten auf Patientenebene. Die Wahl einer geeigneten Zeiteinheit und Einteilung der Zeitachse für die Ableitung eines aussagekräftigen und Patienten-relevanten Hauptendpunkts (*Summary Outcome*) ist ein wesentlicher Aspekt bei der Entwicklung eines biometrischen Konzepts für die Wirksamkeitsschätzung. Die Vorgehensweise wird im Studienprotokoll oder spätestens im SAP vor Entblindung festgelegt. In Abhängigkeit von der Beobachtungsdauer, Annahmen hinsichtlich des erwarteten Behandlungseffekts der experimentellen Therapie, sowie krankheitsspezifischen Überlegungen wie Inzidenz des interessierenden Ereignisses und der intraindividuellen Variabilität über die Zeit oder die Erfassung eventueller zyklischer Schwankungen, wird typischerweise über ein Zeitintervall von einem Monat (30 Tage) oder einer Woche aggregiert, gerechnet ab einem vordefinierten Startpunkt[§]. Eine Alternative zu einer äquidistanten Aufteilung der individuellen Zeitachse für die Ableitung der Anzahl der Attacken und Anzahl der Beobachtungstage pro Zeiteinheit wäre eine Zeiteinteilung in Abhängigkeit von den im Studienprotokoll geplanten Visiten (siehe z. B. Vorgehensweise bei Bunouf *et al.*, 2012). Ein Nachteil dieser Strategie ist unter anderem ein möglicher Selektionsbias aufgrund informativer Visitenzeitpunkte, wenn Visiten tatsächlich nicht im vordefinierten Zeitfenster stattfinden oder komplett entfallen.

2.3.2 Spezifikation eines multiplikativen Modells

Durch die Ableitung des Summary Outcomes ergeben sich natürlicherweise longitudinale Zähldaten, mit vollständigem Response-Profil über die Zeit bei Patienten mit komplettem Follow-up (Completer) bzw. unvollständigem Follow-up bei vorzeitigen Studienabbrechern (Analyse-Dropouts) oder fehlender Tagebuchdokumentation zu bestimmten Zeitpunkten.

[§]Als patientenindividueller Startpunkt (Tag 1) wird in der Regel der Zeitpunkt der Randomisierung oder Start der Behandlung gewählt.

Bei RCTs mit Repeated Measures Design zum Nachweis der Überlegenheit einer experimentellen Therapie im Vergleich zu Placebo oder einer Standardtherapie ist der auf dem primären Studienziel basierende Efficacy-Endpunkt häufig ein bestimmter Zeitpunkt oder Zeitintervall (Assessment Periode) am Ende einer länger andauernden Intervention. Eine konfirmatorische Hauptanalyse nach dem Intention-to-Treat Prinzip mit einfachen nonparametrischen Tests für den Gruppenvergleich oder einer Analysis of Covariance (ANCOVA) mit Adjustierung für den entsprechenden Baseline-Wert bei normalverteiltem Outcome zur Schätzung des Treatment-Effekts zu einem vordefinierten Zeitpunkt setzt unter anderem gleiche Informationszeiten für alle Patienten der Analysepopulation voraus und ignoriert die durch verschiedene Dropout Mechanismen verursachte Komplexität der Verlaufsdaten. Solche Complete-Case Analysen liefern valide, unverzerrte Schätzergebnisse nur unter der sehr restriktiven Annahme MCAR, bei zusätzlichem Powerverlust aufgrund fehlender Daten (vgl. z. B. Ashbeck & Bell, 2016). Ist der Vergleich zwischen den Behandlungsgruppen zu mehreren Zeitpunkten von klinischem Interesse, erfordert dies zudem eine Auswertungsstrategie, welche das multiple Testproblem adäquat berücksichtigt. Auch aus klinischer Sicht ist die Abbildung von patientenspezifischen Response-Profilen im Studienverlauf in den einzelnen Behandlungsgruppen von besonderem Interesse.

Im Vergleich zu Single-time-point Analysestrategien erlauben parametrische, longitudinale Ansätze aus der Klasse der *Mixed-Effects Modelle* (Random Effects Modell, Mixed Model for Repeated Measures (MMRM)) eine effiziente Schätzung des Treatment-Effekts unter Ausnutzung der vollen Information der beobachteten Daten aller Patienten der Analysepopulation unabhängig vom Dropout-Status (Carpenter & Kenward, 2007). Des Weiteren lassen sich Informationen ableiten über die patientenspezifische Veränderung des primären quantitativen Zielkriteriums (Attackenrate) im Verlauf der Beobachtung, und somit Schätzer für die mittlere Geschwindigkeit des sich einstellenden Behandlungseffekts (‘*Speed of Effect*’, vgl. CHMP (2007)) in den untersuchten Behandlungsgruppen gewinnen. Ein Likelihood-basiertes Modell mit *konditionaler* Sichtweise liefert im Gegensatz zu marginalen bzw. populationsspezifischen Ansätzen (z. B. Generalized Estimating Equations (GEE)) valide Schätzungen unter Missing Completely at Random (MCAR) sowie der weniger restriktiven Annahme Missing at Random (MAR) (Little & Rubin, 2002), und wird u. a. in der Efficacy Guideline der EMA zu fehlenden Werten im regulatorischen Setting konfirmatorischer RCTs als geeignete Primäranalyse vorgeschlagen, welche ohne explizite Imputationsverfahren auskommt (CHMP, 2010; Mallinckrodt *et al.*, 2003; Molenberghs *et al.*, 2004). Bei longitudinalen Zähldaten erfolgt die Wirksamkeitsschätzung analog unter Verwendung von Generalisierten linearen gemischten Modellen (GLMMs) für nicht-normalverteilte Zielgrößen (Breslow & Clayton, 1993).
Für theoretische Details zu Gemischten Modellen sei z. B. auf Laird & Ware (1982), Diggle *et al.* (2003) oder Verbeke & Molenberghs (2005) verwiesen.

Notation Sei y_{ij} das longitudinale Summary Outcome (Anzahl der Attacken) für Patient i ($i = 1, \ldots, N$) pro Zeiteinheit j, t_{ij} ($j = 1, \ldots, n_i$) der Messzeitpunkt nach Randomisierung, N die Gesamtzahl der randomisierten Patienten in der Studie, und n_i die Anzahl der Follow-up Perioden pro Patient. Bei einer äquidistanten Zeiteinteilung (vgl.

Abschnitt 2.3.1) und diskreter Zeitvariable gilt $t_{ij} \equiv t = 1, 2, \ldots, n_i$, wobei n_i das letzte Zeitintervall unter Beobachtung bezogen auf das primäre Zielkriterium $y_i(t)$ darstellt. Die Offset-Variable $\log(d_i(t))$ sei die logarithmierte Anzahl bewertbarer Tage im Intervall t und kann als Maß für die Exposition eines Patienten i im Intervall t interpretiert werden, welche über die Zeit variierten kann. Bei einer RCT mit zwei Behandlungsgruppen (Treatment-Indikator: $x_i = 0$ Placebo, $x_i = 1$ experimentelle Behandlung) und Annahme eines konstant *linearen Zeittrends* auf der Skala des linearen Prädiktors kann folgende Mittelwertstruktur als Ausgangspunkt für die Spezifikation eines *saturierten* gemischten Modells zugrundegelegt werden:

$$\eta_i(t) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) \cdot t + \beta_2\, x_i + \beta_3\, x_i \cdot t + \log\left(d_i(t)\right) ,$$

wobei $\boldsymbol{b}_i = (b_{0i}, b_{1i})^\mathsf{T} \sim \mathcal{N}_2(\mathbf{0}, \boldsymbol{\Sigma})$ der bivariat normalverteilte Vektor für die i-ten patientenspezifischen, miteinander korrelierten Random Effects – Intercept und Slope bezogen auf die Zeit – mit Erwartungswert Null und Varianz-Kovarianzmatrix $\boldsymbol{\Sigma}$.
Sei $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)^\mathsf{T}$ der populationsspezifische Parametervektor für die festen unbekannten Effekte, mit Intercept (β_0), Haupteffekt für die Zeit (β_1 Faktor für die mittlere Veränderung unter Placebo), Haupteffekt für die Behandlungsgruppe (β_2), sowie der interessierende Parameter β_3 (Interaktionseffekt Treatment $\times$ Zeit) zur Quantifizierung des globalen Treatment-Effekts über die Zeit, interpretierbar als Unterschied in der Steigung ('Speed of Effect') unter der experimentellen Behandlung im Vergleich zu Placebo. Bei einem modellbasierten Analyseansatz kann die primäre Nullhypothese H_0 dann wie folgt formuliert werden:

$$H_0 : \text{Treatment} \times \text{Zeit -Interaktion} = 0.$$

Gängige Verteilungsannahmen bei Zähldaten Ausgangspunkt für eine prospektive Modellbildung bei longitudinalen Zähldaten ist die Annahme einer geeigneten Verteilung für $\mathbf{y}_i$. Hierbei dient die Poisson-Verteilung mit der restriktiven Annahme der Equidispersion (d. h. Varianz gleich Erwartungswert) als Referenzverteilung, und es gilt der *loglineare* Zusammenhang $\log(\mu_i(t)) = \eta_i(t)$, mit $\mu_i(t)$ die erwartete Attackenanzahl im Intervall t, und $\mu_i(t)/d_i(t)$ die erwartete Attackenrate pro Zeiteinheit. In der Praxis ist häufig die Variabilität von y_{ij} höher als unter der Poisson-Verteilung erwartet. Ursachen für *Überdispersion* ist z. B. 1.) unbeobachtete interindividuelle Heterogenität aufgrund fehlender prädiktiver Faktoren (Baseline-Kovariablen), 2.) die Korrelation zwischen den Beobachtungen (hierarchische bzw. longitudinale Datenstruktur), oder 3.) Zero-Inflation, d. h. ein überproportionaler Anteil an Nullen, welcher höher ist als unter der Poisson-Verteilung erwartet (Lambert, 1992). Eine Fehlspezifikation durch Nicht-Berücksichtigung der Überdispersion kann zu verzerrten Parameterschätzungen und einer Unterschätzung der entsprechenden Standardfehler, und somit zu falschen Schlussfolgerungen hinsichtlich der untersuchten Kovariablen-Effekte $\boldsymbol{\beta}$ führen.

Um in der Zähldatensituation mit Überdispersion umzugehen, werden in der Literatur verschiedene flexible Modellerweiterungen vorgeschlagen. Für ein GLMM basierend auf der Negativ Binomialverteilung (NB) mit zusätzlichem Dispersionsparameter k (üblicherweise

als identisch in den Behandlungsgruppen angenommen[¶]) wird in der klassischen Parametrisierung ein quadratischer Zusammenhang zwischen Erwartungswert und Varianz zugrundegelegt, $\mathrm{Var}\,(y_i) = \mu_i + \mu_i^2/k$, vgl. die Typ 2-Parametrisierung in Hilbe (2011). In Abbildung 2.2 ist für den gepoolten Datensatz der BEMED-Studie der Zusammenhang zwischen Mittelwert und Varianz der Attackeninzidenz auf Patientenebene dargestellt. Ohne Kenntnis der Behandlungsgruppe finden sich bei dieser einfachen graphischen Darstellung bereits Hinweise für unspezifische Überdispersion, und eine Poisson-Regression erscheint nicht adäquat.

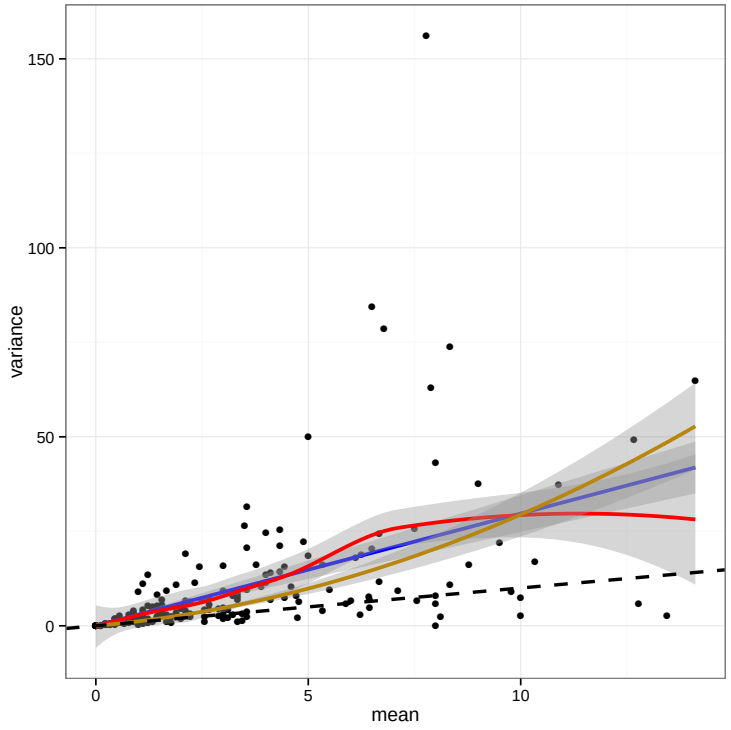

Abbildung 2.2: Verblindeter Attackendatensatz der BEMED-Studie (Full Analysis Set, $N = 213$ Patienten). Zusammenhang zwischen Erwartungswert und Varianz auf Patientenebene (mit 95% Konfidenzband): schwarze gestrichelte Linie mit Steigung=1: Poisson [$\mathrm{Var}\,(y_i) = \mu_i$]; orangefarbene Kurve: quadratischer Zusammenhang bei Verwendung der üblichen Parametrisierung der Negativ Binomialverteilung (NB, Typ 2-Parametrisierung nach Hilbe (2011)); blaue Linie: linearer Zusammenhang (NB, Typ 1-Parametrisierung (Quasi-Poisson) nach Hilbe (2011)); rot: nonparametrischer Zusammenhang (Scatterplot-Smoother).

In manchen Datensituationen eignet sich die Normalverteilungsannahme nach einer streng monotonen, asymptotisch varianzstabilisierenden Transformation einer Responsevariable y, gegeben eine bestimmte Mittelwertstruktur (Tibshirani, 1988; Hastie & Tibshirani,

[¶]Die Annahme eines über die Behandlungsgruppen homogenen Dispersionsparameters ist eine relative starke Modellannahme, welche kaum verifizierbar ist. Liegt in Wahrheit eine Gruppenspezifische Überdispersion vor, können Inferenzmethoden mit gepooltem Dispersionsparameter verzerrte Schätzergebnisse liefern.

1990, Kap. 7.4). Nach Adrion & Mansmann (2012, Appendix A1) resultiert für negativ binomial-verteiltes y die arcus-sinus-hyperbolicus Transformation (Jeffrey, 2000), definiert über $\tilde{y} := \text{arcsinh}(y) = \log(y + \sqrt{y^2 + 1})$, und führt zu asymptotisch normalverteilten Residuen auf der transformierten Skala $\tilde{y}$ mit annähernd konstanter Varianz (Homoskedastizität). Homoskedastische Daten erlauben die Anwendung analytisch einfacherer Inferenzmethoden. Sofern methodisch gerechtfertigt vereinfacht diese Transformation vor allem im frequentistischen Kontext eine Parameterschätzung erheblich und ermöglicht eine Likelihood-basierte Modellierung für normalverteilte hierarchische Daten. Nachteile dieser Response-Transformation für negativ binomial-verteilte Zähldaten ergeben sich bei der Interpretierbarkeit und bei der Modellselektion aufgrund der geänderten Skalierung (Modellierung von $\tilde{y}$ anstatt einer loglinearen Modellierung von y auf der Originalskala).

In bestimmten komplexeren Datensituationen ist ein Modellierungsansatz basierend auf der Annahme einer Negativ Binomial-Verteilung nicht effizient: Ist die zusätzliche Variabilität verursacht durch einen überproportionalen Anteil an Nullwerten relativ zur zugrundeliegenden datengenerierende Verteilung (Poisson oder Negativ Binomial) eignen sich Modellerweiterungen mit *Zero-Inflation* (ZI) (Lambert, 1992; Warton, 2005) für eine möglichst unverzerrte Schätzung der interessierenden Wirksamkeits-Parameter. Beispielsweise ermöglicht ein GLMM mit Annahme einer Zero-inflated Negativ Binomial-Verteilung den Umgang mit Daten bei Überdispersion, sofern die zusätzliche Variabilität verursacht ist durch unbeobachtete Heterogenität und zusätzliche Nullen, die kategorisierbar sind als 'strukturelle' und zufällig auftretende Nullen ('sampling zeros') (Yau *et al.*, 2003). Zero-inflated Modelle basieren auf einer Mischverteilung von Zähldaten inklusive Nullen ('at-risk group', 'not always zero-group') und einem Punktmaß bei Null ('not at-risk group', 'always zero-group'). Übertragen auf die Daten der BEMED-Studie wären Menière-Patienten der 'not at-risk group' und einer Attackeninzidenz gleich Null im Beobachtungszeitraum entweder Patienten in einer inaktiven Phase ihrer Erkrankung, oder Therapie-Responder, d. h. Patienten, die auf eine Intervention ansprechen, welche die Attackenrate auf Null senkt.

Für eine Übersicht zur Klasse der sogenannten Two-part Mixture Modelle für Zähldaten mit Zero-Modifikation sei z. B. auf Neelon *et al.* (2016) oder Zuur *et al.* (2009) verwiesen. Diese Referenzen enthalten auch methodische Details zum Hurdle-Modell. Im Gegensatz zum Zero-inflated Modell basiert dieses Mixture-Modell auf zwei Komponenten, einer bei Null trunkierten Verteilung (Poisson oder NB) für streng positive Zähldaten, sowie auf einem Punktmaß bei Null, und ermöglicht den Umgang mit Zero-Deflation (weniger Nullen als unter der datengenerierenden Poisson- bzw. NB-Verteilung erwartet). Wesentliche Annahme bei einem Hurdle-Modell ist die Aufteilung der Patienten in eine Gruppe mit Null Ereignissen (Attackeninzidenz) versus eine Gruppe mit einer streng positiven Anzahl an Ereignissen. Inhaltliche Überlegungen sind Voraussetzung dafür, ob diese Annahme einer kompletten Separation für die untersuchte Fragestellung und Erkrankung *a priori* gerechtfertigt erscheint und für einen möglichen Mixed-Effects Modellierungsansatz in Betracht zu ziehen ist.

3 Zusammenfassende Bewertung und Vorstellung der Beiträge

Diese kumulative Dissertation besteht aus zwei Publikationen mit Supplements, sowie einem Appendix mit Bezug zu beiden Publikationen.

Der erste Artikel beschreibt die statistische Methodik bei der Planung der Hauptanalyse einer konfirmatorischen verblindeten Phase III-Studie mit longitudinalen Attackendaten von Morbus Menière-Patienten aus der Sicht des verantwortlichen Biometrikers. Es werden Lösungsansätze zur transparenten Entscheidungsfindung bei der Entwicklung eines *Statistischen Analyseplans* (SAP) aufgezeigt mit dem Ziel, eine modellbasierte Auswertungsstrategie für den primären Endpunkt bei longitudinalen Zähldaten zu präspezifizieren. Die in diesem methodischen Artikel behandelte Problemstellung, relevant in der Projektphase der endgültigen Finalisierung des SAPs, somit vor der Sperrung der Datenbank mit nachfolgender Entblindung, wird am Beispiel der multizentrischen BEMED-Studie aufgezeigt. Der zweite anwendungsorientierte Artikel präsentiert und diskutiert umfassend die Efficacy- und Safety-Resultate dieser konfirmatorischen RCT.

Im Folgenden werden die Inhalte der beiden verwendeten Fachartikel sowie des Anhangs kurz vorgestellt, und jeweils der Beitrag zu den einzelnen Veröffentlichungen dargelegt.

Paper I: Bayesian model selection techniques as decision support for shaping a statistical analysis plan of a clinical trial: An example from a vertigo phase III study with longitudinal count data as primary endpoint

Dieser methodisch orientierte Artikel behandelt biometrische Prinzipien und Strategien bei der Entwicklung eines SAP für eine konfirmatorische, verblindete, randomisierte, kontrollierte klinische Studie (Phase IIb oder Phase III) mit longitudinalen Zähldaten als primäres Efficacy Outcome. Die Wahl eines adäquaten parametrischen Modells für die primäre Wirksamkeitsanalyse oder für zugehörige *Sensitivitätsanalysen* oder *Zusatzanalysen** (Moher *et al.*, 2010; Schulz *et al.*, 2010) erfordert Entscheidungen, die *a priori*, d. h.

*im CONSORT-Statement hierfür verwendete Begriffe: 'additional' bzw. 'ancillary analyses'

nicht datengesteuert, getroffen und im SAP festgelegt werden müssen. Der Artikel konzentriert sich auf zwei wichtige Aspekte bei der Spezifikation eines generalisierten linearen gemischten Modells (GLMM), nämlich

1. Annahmen bezüglich der Verteilungsstruktur:
   Poisson- vs. Negativ Binomial-Verteilung zur Berücksichtigung von Überdispersion; Erweiterungen aufgrund von potentieller Zero-Inflation; Normalverteilung nach asymptotisch varianzstabilisierender Transformation ("Symmetrisierung") des primären Efficacy Outcomes (Hastie & Tibshirani, 1990)
2. Annahmen bezüglich der Varianz-Kovarianzstruktur:
   patientenspezifischer Intercept bzw. patientenspezifische Steigung.

Am Beispiel der BEMED-Studie wird die Komplexität der Modellfindung für die Primäranalyse der Attackendaten aus der Sicht des Studienstatistikers aufgezeigt: Anhand von longitudinalen Attackendaten der explorativen Betahistin-Studie[†] (Strupp *et al.*, 2008a), einer Anwendungsbeobachtung, deren Endpunktdaten in wesentlichen Strukturmerkmalen (inkl. gleiche Indikation, gleiche Therapie, gleiches Patientenkollektiv) mit denen der Hauptstudie vergleichbar sind, wird im Bayesianischen Setting eine *vorhersageorientierte* Modellwahl im Rahmen einer Complete-Case-Analyse[‡] durchgeführt. Diese Strategie wird in der englischsprachigen Fachliteratur auch als '*informed model choice*' bezeichnet.

Der Artikel diskutiert verschiedene, universell einsetzbare Bayesianische Tools zur Selektion und Validierung nicht genesteter longitudinaler Modelle, wie z. B. das Deviance Information Criterion (DIC), sowie Leave-one-out kreuzvalidierte Kriterien (LOOCV), welche auf der posteriori prädiktiven Verteilung basieren. Dazu gehören u. a. die Conditional Predictive Ordinate (CPO), Logarithmische Scores als Beispiel für eine (strikt korrekte) Bewertungsregel (Proper Scoring Rules) zur Beurteilung der Prognosegüte eines hierarchischen Modells, sowie die Probability Integral Transform (PIT) (Czado *et al.*, 2009; Gneiting & Raftery, 2007). Darüber hinaus wird die praktische Umsetzung mit der von Rue *et al.* (2009) entwickelten INLA Methodik demonstriert, einem flexiblen und effizienten Verfahren zur deterministischen approximativen Bayes-Inferenz für sog. latente Gauß-Modelle mittels Integrated Nested Laplace Approximation (siehe z. B. Fong *et al.*, 2010; Martino & Rue, 2010a,b; Rue *et al.*, 2013). INLA stellt einen alternativen Ansatz zur klassischen Bayes-Inferenz mittels Sampling-basierter Verfahren wie MCMC (Markov Chain Monte Carlo) dar. Im Gegensatz zu MCMC erfolgt bei INLA kein Sampling aus der Posteriori-Verteilung, stattdessen wird die Posteriori-Verteilung in 'geschlossener Form' approximiert. Somit sind die bei komplexeren Bayesianischen hierarchischen Modellen unter MCMC häufiger auftretenden Probleme fehlender Konvergenz oder schlechter

[†]Studiendesign: 2-armige unverblindete, nicht placebokontrollierte, nicht randomisierte, monozentrische Anwendungsbeobachtung zum Vergleich der Wirksamkeit von Betahistin in niedriger vs. höherer Dosierung; Fallzahl: $N = 112$ Patienten (vgl. Seite 7)

[‡]Die Modellwahl erfolgte anhand von Daten der 112 Studienpatienten, für die vollständige Attackenverläufe (d. h. dokumentiert über den gesamten Studienverlauf) vorlagen. Daten von 16 Patienten mit Therapie- und/oder Beobachtungsabbruch waren für dieses Projekt nicht verfügbar. Somit handelt es sich um eine Analyse, welcher implizit die Missing Completely at Random (MCAR)-Annahme zugrunde lag (National Research Council, 2010; Carpenter & Kenward, 2007; Little & Rubin, 2002).

Mixing-Eigenschaften, mit denen der Anwender gerade bei nicht-normalverteiltem Outcome (z. B. Zähldaten) häufiger konfrontiert ist, nicht relevant.
Dieses Bayesianische Instrumentarium erweist sich als in der Praxis geeignet, um mithilfe von externen Daten einer Pilot-, Feasibility-Studie oder Anwendungsbeobachtung (im Artikel: Betahistin-Vorstudie) ein GLMM zu spezifizieren, welches Güteeigenschaften wie Robustheit, Einfachheit und prädiktive Performance erfüllt (Gelfond *et al.*, 2011). Letztendlich kann auf diese Art eine transparente und begründbare, somit eine *informierte* Entscheidung für ein bestimmtes Mixed Effects Modell (im frequentistischen oder Bayesianischen Setting) erfolgen, welches im SAP einer verblindeten konfirmatorischen RCT höherer Evidenz (im Artikel: Phase III BEMED-Studie) vorab spezifiziert wird, ohne Kenntnis der Studiendaten inklusive Therapiezuteilung. Mit dieser Vorgehensweise, welche eine datengesteuerte Modellwahl ausschließt, kann die konfirmatorische Validität der modellbasierten Analyse für die primäre Fragestellung einer RCT gewährleistet werden.

Des Weiteren enthält der Artikel eine Simulationsstudie, in der die Performance des DIC sowie des Logarithmischen Scores untersucht wurde. Hierbei wurden longitudinale Zähldaten unter der Negativ Binomial-Verteilung generiert – bei Variation des Grades an Überdispersion (quantifizierbar anhand des Überdispersionsparameters k) und der Fallzahl – und die diskriminatorische Power dieser beiden Bayesianischen Tools unter der Annahme möglicher Modellalternativen aus der Klasse der gemischten Modelle für Zähldaten (Negativ Binomial, Poisson (mit und ohne Zero-Inflation), Normal-Verteilung nach varianzstabilisierender arcus-sinus-hyperbolicus Transformation) untersucht.
Das zugehörige Web Supplement skizziert die Implementation der konkurrierenden Modelle und Bayesianischen Tools zur Modellevaluation in `R-INLA`.

Die Doktorandin war Erstautorin dieses Artikels und damit hauptverantwortlich für die Ausarbeitung des gesamten Manuskripts, die Durchführung der statistischen Analysen des realen Datenbeispiels sowie der Simulationsstudie in INLA, und führte sämtliche Programmierarbeiten in R selbständig durch.

Paper II: Primärpublikation der BEMED-Studie – Hauptergebnisse zur Wirksamkeit und Sicherheit

Dieser Artikel stellt das Publikationsmanuskript der BEMED-Studie dar, dessen Aufbau und Inhalt sich an der CONSORT-PRO Reporting Guideline (2013) für RCTs mit patientenorientierten Efficacy Outcomes orientiert.

Es konnte kein positiver Effekt der Betahistin-Therapie im Vergleich zu Placebo nachgewiesen werden, Betahistin in der untersuchten Tagesdosis von 48 mg (Standarddosis) bzw. 144 mg (experimentelle Hochdosis) ist nicht wirksamer als Placebo. Die Studiendaten liefern keine Evidenz für einen Unterschied in der Attackenrate zwischen den drei Behandlungsarmen. Es konnte kein Behandlungseffekt von Betahistin (Standarddosis, Hochdosis) im Vergleich zur Placebo-Intervention nachgewiesen werden ($P_{\text{global}} = 0.759$,

Likelihood Ratio-Test): Für das Full Analysis Set nahm die monatliche Attackenrate innerhalb der 9-monatigen Behandlungsdauer in allen drei Gruppen um den Faktor 0.758 (95% KI: 0.705; 0.816), $P < 0.001$, ab. Im Vergleich zu Placebo ergab sich für die Standarddosis Betahistin-Gruppe ein Rate Ratio von 1.036 (0.942; 1.140), für die Hochdosis Gruppe ein Rate Ratio von 1.012 (0.919; 1.114). Die populationsbasierte (d. h. marginale[§]) mittlere Attackenrate pro Monat innerhalb der 90-tägigen Assessment-Periode (Monat 7 bis 9) war 2.722 (1.304; 6.309) unter der Placebo-Intervention, 3.204 (1.345; 7.929) unter der Standarddosis, und 3.258 (1.685; 7.266) unter der experimentellen Hochdosis Betahistin. Aufgrund einer fehlenden Kontroll-Gruppe, welche keinerlei Intervention erhält ('no-treatment' Arm), konnte in der BEMED-Studie nicht differenziert werden zwischen einem wahren Placebo-Effekt und anderen unspezifischen Effekten wie natürlicher (fluktuierender) Verlauf der Attackenrate, Spontanremission, zeitlichen Effekten oder Regression-to-the-mean [¶] (Enck *et al.*, 2013; Hamill, 2006).

Das zugehörige Web Supplement enthält u. a. weitere methodische Details zu den im SAP präspezifizierten Sensitivitäts- sowie Zusatzanalysen zur Efficacy- und Effectiveness-Fragestellung, um die Robustheit des Hauptergebnisses aus statistischer Sicht zu demonstrieren, sowie Definitionen sekundärer Efficacy Outcomes.

Die Doktorandin war als Erstautorin hauptverantwortlich für die Erstellung des gesamten Manuskriptentwurfs inklusive Abschnitt zur klinisch-biometrischen Diskussion, sowie für die Bearbeitung der nachfolgenden Revision im Rahmen des Review-Prozesses. Des Weiteren war sie zusammen mit Frau Dr.med. Carolin Simone Fischer Mitglied im zentralen *Endpoint Assessment Committee*, implementiert am Studienzentrum des Sponsors, dessen Aufgabe die verblindete, vollständige Evaluation aller BEMED-Tagebücher darstellte mit dem Ziel einer möglichst objektiven und standardisierten Ableitung der Efficacy-Daten anhand der patientenberichteten vestibulären Symptome (Details siehe zugehörige SOP von Fischer *et al.*, 2014). Sie war als Biometrikerin der BEMED-Studie hauptverantwortlich für die Durchführung sämtlicher statistischer Analysen und deren Interpretation, sowie für die Erstellung des Statistical Reports gemäß SAP, welcher diesem Artikel zugrundeliegt.

APPENDIX:
Statistischer Analyseplan für die BEMED-Studie

Den Anhang dieser Dissertation bildet der SAP für die BEMED-Studie. Dieser SAP enthält eine methodische Beschreibung des Studiendesigns und -ablaufs, definiert primäre und sekundäre Studienziele sowie die zugehörigen Efficacy- und Safety-Endpunkte. Darüber hinaus werden die Analysepopulationen für Wirksamkeits- und Sicherheitsanalysen (Full Analysis Set, Per Protocol Set; Safety Set Sample) sowie deren Herleitung anhand

[§]Diese marginalen Schätzer für die Attackeninzidenz wurden im Bayesianischen Setting (MCMC) von den aus dem Negativ Binomial Mixed Model resultierenden konditionalen Schätzern abgeleitet.

[¶]Der Studieneinschluss erfolgte in einer aktiven Krankheitsphase, bei einer bestimmten Symptomschwere, welche definiert wurde über bestimmte Einschlusskriterien.

studienspezifischer Kriterien festgelegt. Die im Prüfplan lediglich skizzierten Auswertungsmethoden werden im SAP präspezifiziert und detailliert beschrieben inklusive der zugrundeliegenden (prüfbaren) Annahmen, insbesondere für die konfirmatorische Hauptanalyse, zur Sicherstellung der internen Validität und zur Minimierung von Analyse-Bias.

Der komplexe Prozess des Tagebuch-Assessments vor Entblindung mit dem Ziel der Operationalisierung und standardisierten Ableitung des primären Efficacy Outcome, d. h. der Anzahl der Menière-assoziierten Schwindelattacken pro Zeiteinheit, durch eine regelbasierte klinische Bewertung und Klassifikation der in den Patienten-Tagebüchern auf täglicher Basis dokumentierten Schwindelsymptome (*Rohdaten*) wird in der SOP von Fischer, Adrion & Strupp (2014) umfassend beschrieben (vgl. Appendix I des SAP). Dieses studien- und krankheitsspezifische Consensus-Dokument ist einer der wesentlichen Bestandteile des SAP und wurde entwickelt, um vor Entblindung eine möglichst valide und reliable Ableitung der Attackendaten für die Hauptfragestellung der BEMED-Studie zu gewährleisten und Kriterien u. a. für den Umgang mit unvollständiger oder fehlerhafter Patientendokumentation festzulegen.

Die Doktorandin war Autorin des SAP. Die Spezifizierung der Efficacy- und Safety-Analysen erfolgte verblindet in Unkenntnis des Behandlungscodes (Placebo vs. Standarddosis vs. Hochdosis Betahistin). Der SAP wurde nach dem Blinded Data Review und vor der Entblindung der offiziellen Studiendatenbank Ende Juli 2014 finalisiert und durch die Sponsor Delegated Person genehmigt.

Literaturverzeichnis

AAO-HNS. American Academy of Otolaryngology – Head and Neck Surgery Foundation. Committee on Hearing and Equilibrium. Guidelines for the diagnosis and evaluation of therapy in Menière's disease. *Otolaryngology – Head and Neck Surgery* 1995; **113**(3):181–185.

Adrion C, Mansmann U. Bayesian model selection techniques as decision support for shaping a statistical analysis plan of a clinical trial: An example from a vertigo phase III study with longitudinal count data as primary endpoint. *BMC Medical Research Methodology* 2012; **12**(1):137.

Ashbeck EL, Bell ML. Single time point comparisons in longitudinal randomized controlled trials: power and bias in the presence of missing data. *BMC Medical Research Methodology* 2016; **16**(1):1–8.

Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient reported outcomes. *Statistical Methods in Medical Research* 2014; **23**(5):440–59.

Brandt T, Dieterich M, Strupp M. *Vertigo: Leitsymptom Schwindel*. Steinkopff, 2004.

Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* 1993; **88**(421):9–25.

Bunouf P, Grouin JM, Molenberghs G. Analysis of an incomplete binary outcome derived from frequently recorded longitudinal continuous data: application to daily pain evaluation. *Statistics in Medicine* 2012; **31**(15):1554–1571.

Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD, CONSORT PRO Group. Reporting of patient-reported outcomes in randomized trials: The CONSORT PRO Extension. *JAMA* 2013; **309**(8):814–822.

Cappelleri JC, Bushmakin AG. Interpretation of patient-reported outcomes. *Statistical Methods in Medical Research* 2014; **23**(5):460–83.

Carpenter JR, Kenward MG. *Missing data in randomised controlled trials – a practical guide*. National Institute for Health Research, Birmingham, 2007. Publication RM03/JH17/MK. Available at `http://researchonline.lshtm.ac.uk/4018500/1/rm04_jh17_mk.pdf`. Last accessed March 15, 2018.

Chan AW, Tetzlaff JM, Altman DG, Laupacis A, Gøtzsche PC, Krleža-Jerić K, Hróbjartsson A, Mann H, Dickersin K, Berlin JA, Doré CJ, Parulekar WR, Summerskill WS, Groves T, Schulz KF, Sox HC, Rockhold FW, Rennie D, Moher D. SPIRIT 2013 Statement: Defining Standard Protocol Items for Clinical Trials. *Annals of Internal Medicine* 2013a; **158**(3):200–207.

Chan AW, Tetzlaff JM, Gøtzsche PC, Altman DG, Mann H, Berlin JA, Dickersin K, Hróbjartsson A, Schulz KF, Parulekar WR, Krleža-Jerić K, Laupacis A, Moher D. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ* 2013b; **346**:e7586.

CHMP. *Reflection paper on the regulatory guidance for the use of health-related quality of life (HRQL) measures in the evaluation of medicinal products*. European Medicines Agency, Committee for Medicinal Products for Human Use (CHMP), London, UK, 2005. URL `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003637.pdf`. EMEA/CHMP/EWP139391/2004.

CHMP. *Guideline on clinical investigation of medicinal products for the treatment of migraine*. European Medicines Agency, Committee for Medicinal Products for Human Use (CHMP), London, UK, 2007. URL `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003481.pdf`. CPMP/EWP/788/2001 Rev. 1. Date for coming into effect: 31 July 2007.

CHMP. *Guideline on Missing Data in Confirmatory Clinical Trials*. European Medicines Agency, Committee for Medicinal Products for Human Use (CHMP), London, UK, 2010. URL `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2010/09/WC500096793.pdf`. EMA/CPMP/EWP/1776/99 Rev. 1. Date for coming into effect: 1 January 2011.

CHMP. *Draft Guideline on the investigation of subgroups in confirmatory clinical trials*. European Medicines Agency, Committee for Medicinal Products for Human Use (CHMP), London, UK, 2014a. URL `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/02/WC500160523.pdf`. EMA/CHMP/539146/2013. First published 03/02/2014.

CHMP. *Draft reflection paper on the use of patient reported outcome (PRO) measures in oncology studies*. European Medicines Agency, Committee for Medicinal Products for Human Use (CHMP), London, UK, 2014b. URL `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2014/06/WC500168852.pdf`. EMA/CHMP/292464/2014. First published 17/06/2014.

CHMP. *Guideline on adjustment for baseline covariates in clinical trials*. European Medicines Agency, Committee for Medicinal Products for Human Use (CHMP), London, UK, 2015. URL `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2015/03/WC500184923.pdf`. EMA/CHMP/295050/2013. Adopted and first published 27/03/2015.

Czado C, Gneiting T, Held L. Predictive model assessment for count data. *Biometrics* 2009; **65**(4):1254–1261.

Della PC, Guidetti G, Eandi M. Betahistine in the treatment of vertiginous syndromes: a meta-analysis. *Acta Otorhinolaryngologica Italica* 2006; **26**(4):208–215.

Diggle P, Heagerty P, Liang KY, Zeger S. *Analysis of longitudinal data*. Oxford University Press, Oxford, 2nd ed., 2003.

Enck P, Bingel U, Schedlowski M, Rief W. The placebo response in medicine: minimize, maximize or personalize? *Nature Reviews Drug Discovery* 2013; **12**(3):191–204.

Fairclough DL. Patient reported outcomes as endpoints in medical research. *Statistical Methods in Medical Research* 2004; **13**(2):115–138.

FDA. *Guidance for Industry – Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*. Food and Drug Administration, U.S. Department of Health and Human Services, Center for Drug Evaluation and Research, 2009. URL `https://www.fda.gov/downloads/drugs/guidances/ucm193282.pdf`. Last accessed November 30, 2016.

Fischer C, Adrion C, Strupp M. SOP "Diary Assessment": Verblindete Attackenbewertung der Patienten-Tagebücher der *BEMED* Studie – unveröffentlichte Standard Operating Procedure, Version 1.2, Mai, 2014. Offizieller Anhang des Statistischen Analyseplans der *BEMED* Studie.

Fong E, Li C, Aslakson R, Agrawal Y. Systematic Review of Patient-Reported Outcome Measures in Clinical Vestibular Research. *Archives of Physical Medicine and Rehabilitation* 2015; **96**(2):357–365.

Fong Y, Rue H, Wakefield J. Bayesian inference for generalized linear mixed models. *Biostatistics* 2010; **11**(3):397–412.

Gater A, Coon CD, Nelsen LM, Girman C. Unique challenges in development, psychometric evaluation, and interpretation of daily and event diaries as endpoints in clinical trials. *Therapeutic Innovation & Regulatory Science* 2015; **49**(6):813–821.

Gates G, Verrall A. Validation of the Menière's Disease Patient-Oriented Symptom-Severity Index. *Archives of Otolaryngology – Head & Neck Surgery* 2005; **131**(10):863–867.

Gates GA. Clinimetrics of Meniere's Disease. *The Laryngoscope* 2000; **110**(S94):8–11.

Gelfond JAL, Heitman E, Pollock BH, Klugman CM. Principles for the ethical analysis of clinical and translational research. *Statistics in Medicine* 2011; **30**(23):2785–2792.

Gneiting T, Raftery AE. Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association* 2007; **102**(477):359–378.

Green JD, Verrall A, Gates GA. Quality of Life Instruments in Ménière's Disease. *The Laryngoscope* 2007; **117**(9):1622–1628.

Hamill TA. Evaluating treatments for Menière's disease: controversies surrounding placebo control. *Journal of the American Academy of Audiology* 2006; **17**(1):27–37.

Harcourt J, Barraclough K, Bronstein AM. Meniere's disease – clinical review. *BMJ* 2014; **349**:g6544.

Hastie TJ, Tibshirani RJ. *Generalized Additive Models*, *Monographs on Statistics and Applied Probability*, vol. 43. Chapman & Hall/CRC, London, 1990.

Hilbe JM. *Negative binomial regression*. Cambridge Univ. Press, Cambridge, 2011.

Hróbjartsson A, Gøtzsche PC. Placebo interventions for all clinical conditions. *Cochrane Database of Systematic Reviews* 2010; **1**:CD003974. Review.

Hüfner K, Barresi D, Glaser M, Linn J, Adrion C, Mansmann U, Brandt T, Strupp M. Vestibular paroxysmia diagnostic features and medical treatment. *Neurology* 2008; **71**(13):1006–1014.

ICH E9. *ICH Harmonised Tripartite Guideline E9: Note for Guidance on Statistical Principles for Clinical Trials*. International Conference on Harmonisation, E9 Expert Working Group, 1998. URL `http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500002928.pdf`. CPMP/ICH/363/96. Current Step 5 version.

IHS, (International Headache Society Clinical Trials Subcommittee members): Tfelt-Hansen P, Pascual J, Ramadan N, Dahlöf C, D'Amico D, Diener HC, Hansen JM, Lanteri-Minet M, Loder E, McCrory D, Planca-de S, Schwedt T. Guidelines for controlled trials of drugs in migraine: Third edition. A guide for investigators. *Cephalalgia* 2012; **32**(1):6–38.

Izem R, Kammerman LA, Komo S. Statistical challenges in drug approval trials that use patient-reported outcomes. *Statistical Methods in Medical Research* 2014; **23**(5):398–408.

James AL, Burton MJ. Betahistine for Ménière's disease or syndrome. *Cochrane Database of Systematic Reviews* 2001; **1**:CD001873. Review, Assessed as up-to-date: 24 NOV 2010.

James AL, Thorp MA. Menière's disease. *BMJ Clinical Evidence* 2007; **03**(505).

Jeffrey A. *Handbook of Mathematical Formulas and Integrals*. Inverse Trigonometric and Hyperbolic Functions. Academic Press, Orlando, FL, 2nd ed., 2000. (S. 128–144).

Jones B, Kenward MG. *Design and analysis of cross-over trials*. Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, FL, 2014.

Kammerman LA, Grosser S. Statistical considerations in the design, analysis and interpretation of clinical studies that use patient-reported outcomes. *Statistical Methods in Medical Research* 2014; **23**(5):393–397.

Laird NM, Ware JH. Random-effects models for longitudinal data. *Biometrics* 1982; **38**(4):963–974.

Lambert D. Zero-inflated poisson regression, with an application to defects in manufacturing. *Technometrics* 1992; **34**(1):1–14.

Lezius F, Adrion C, Mansmann U, Jahn K, Strupp M. High-dosage betahistine dihydrochloride between 288 and 480 mg/day in patients with severe menière's disease: a case series. *European Archives of Oto-Rhino-Laryngology* 2011; **268**(8):1237–1240.

Little RJA, Rubin DB. *Statistical analysis with missing data*. Wiley Series in Probability and Statistics. Wiley, New York, 2nd ed., 2002.

Mallinckrodt CH, Clark WS, Carroll RJ, Molenberghs G. Assessing Response Profiles from Incomplete Longitudinal Clinical Trial Data Under Regulatory Considerations. *Journal of Biopharmaceutical Statistics* 2003; **13**(2):179–190.

Martino S, Rue H. Case Studies in Bayesian Computation using `INLA`. In: Mantovan P, Secchi P (Editors), *Complex Data Modeling and Computationally Intensive Statistical Methods*, 99–114. Springer Verlag Italia, Milan, 2010a; .

Martino S, Rue H. *Implementing Approximate Bayesian Inference using Integrated Nested Laplace Approximation: a manual for the* `inla` *program*. Department of Mathematical Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway, 2010b. URL `http://www.math.ntnu.no/~hrue/r-inla.org/doc/inla-manual/inla-manual.pdf`. Last accessed November 30, 2016.

Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, Elbourne D, Egger M, Altman DG. CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; **340**:c869. URL `http://www.consort-statement.org`.

Molenberghs G, Thijs H, Jansen I, Beunckens C, Kenward MG, Mallinckrodt C, Carroll RJ. Analyzing incomplete longitudinal clinical trial data. *Biostatistics* 2004; **5**(3):445–464.

Murdin L, Hussain K, Schilder AG. Betahistine for symptoms of vertigo. *Cochrane Database of Systematic Reviews* 2016; **6**:CD010696.

National Research Council. *The Prevention and Treatment of Missing Data in Clinical Trials*. Panel on Handling Missing Data in Clinical Trials. The National Academies Press, Washington, DC, 2010. URL `https://www.nap.edu/catalog/12955/the-prevention-and-treatment-of-missing-data-in-clinical-trials`. Last accessed March 15, 2018.

Nauta JJ. Meta-analysis of clinical studies with betahistine in Menière's disease and vestibular vertigo. *European Archives of Oto-Rhino-Laryngology* 2014; **271**(5):887–897.

Neelon B, O'Malley AJ, Smith VA. Modeling zero-modified count and semicontinuous data in health services research part 1: background and overview. *Statistics in Medicine* 2016; **35**(27):5070–5093.

Neuhauser H. Epidemiology of vertigo. *Curr Opinion Neurol* 2007; **20**(1):40–6.

Neuhauser H. Epidemiologie von Schwindelerkrankungen. *Der Nervenarzt* 2009; **80**(8):887–894.

Perez-Garrigues H, Lopez-Escamez JA, Perez P, Sanz R, Orts M, Marco J, Barona R, Tapia MC, Aran I, Cenjor C, Perez N, Morera C, Ramirez R. Time course of episodes of definitive vertigo in Menière's disease. *Archives of Otolaryngology – Head & Neck Surgery* 2008; **134**(11):1149–1154.

Rieger A, Mansmann U, Maier W, Seitz L, Brandt T, Strupp M, Bayer O. Versorgungssituation von Patienten mit dem Leitsymptom Schwindel. *Gesundheitswesen* 2014; **76**(6):e32–8.

Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2009; **71**(2):319–392.

Rue H, Martino S, Lindgren F, Simpson D, Riebler A, Krainski ET. *R-INLA: Functions which allow to perform full Bayesian analysis of latent Gaussian models using Integrated Nested Laplace Approximation*. Trondheim, Norway, 2013. URL `www.r-inla.org`. R package version 0.0.

Schulz KF, Altman DG, Moher D. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010; **340**:c332.

Senn S. *Cross-over trials in clinical research*. John Wiley & Sons, 2002.

Silberstein S, Tfelt-Hansen P, Dodick D, Limmroth V, Lipton R, Pascual J, Wang S, for the Task Force of the International Headache Society Clinical Trials Subcommittee. Guidelines for controlled trials of prophylactic treatment of chronic migraine in adults. *Cephalalgia* 2008; **28**(5):484–495.

Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient non-compliance with paper diaries. *BMJ* 2002; **324**(7347):1193–1194.

Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient compliance with paper and electronic diaries. *Controlled Clinical Trials* 2003; **24**(2):182–199.

Strupp M, Brandt T. Leitsymptom Schwindel: Diagnose und Therapie. *Dtsch Arztebl Int* 2008; **105**(10):173–180.

Strupp M, Dieterich M, Brandt T. The treatment and natural course of peripheral and central vertigo. *Dtsch Arztebl Int* 2013; **110**(29–30):505–16.

Strupp M, Hupert D, Frenzel C, Wagner J, Hahn A, Jahn K, Zingler VC, Mansmann U, Brandt T. Long-term prophylactic treatment of attacks of vertigo in Menière's disease – comparison of a high with a low dosage of betahistine in an open trial. *Acta oto-laryngologica* 2008a; **128**(5):520–524.

Strupp M, Kalla R, Claassen J, Adrion C, Mansmann U, Klopstock T, Freilinger T, Neugebauer H, Spiegel R, Dichgans M, Lehmann-Horn F, Jurkat-Rott K, Brandt T, Jen J, Jahn K. A randomized trial of 4-aminopyridine in EA2 and related familial episodic ataxias. *Neurology* 2011; **77**(3):269–275.

Strupp M, Zwergal A, Brandt T. Episodische Ataxien. *Akt Neurol* 2008b; **35**(9):435–442.

Tibshirani R. Estimating transformations for regression via additivity and variance stabilization. *Journal of the American Statistical Association* 1988; **83**(402):394–405.

Verbeke G, Molenberghs G. *Models for Discrete Longitudinal Data*. Springer Series in Statistics. Springer, New York, 2005.

Warton DI. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics* 2005; **16**(3):275–289.

Whitehead A, Jones NMB. A meta-analysis of clinical trials involving different classifications of response into ordered categories. *Statistics in Medicine* 1994; **13**(23-24):2503–2515.

Wood L, Egger M, Gluud LL, Schulz KF, Jüni P, Altman DG, Gluud C, Martin RM, Wood AJG, Sterne JAC. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. *BMJ* 2008; **336**:601.

Yau KKW, Wang K, Lee AH. Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros. *Biometrical Journal* 2003; **45**(4):437–452.

Zuur A, Ieno E, Walker N, Saveliev A, Smith G. *Mixed effects models and extensions in ecology with* R. Springer, 2009.

Publikation I - II

PAPER I

Bayesian model selection techniques as decision support for shaping a statistical analysis plan of a clinical trial: An example from a vertigo phase III study with longitudinal count data as primary endpoint

Christine Adrion & Ulrich Mansmann

Paper published in
BMC Medical Research Methodology 2012; 12(1):137
DOI: 10.1186/1471-2288-12-137

PAPER II

Efficacy and safety of betahistine treatment in patients with Meniere's disease: primary results of a long term, multicentre, double blind, randomised, placebo controlled, dose defining trial (BEMED trial)

Christine Adrion, Carolin S. Fischer, Judith Wagner, Robert Gürkov, Ulrich Mansmann, Michael Strupp; On behalf of the *BEMED* study group

Paper published in *BMJ* 2016; 352:h6816
DOI: 10.1136/bmj.h6816

Web appendix 1: Vertigo diary template
Web appendix 2: Supplementary materials
Web appendix 3: Clinical trial protocol

APPENDIX

Statistical analysis plan for the BEMED trial: a multicenter, double-blind, randomized, placebo-controlled trial on betahistine for the treatment of Menière's disease

Christine Adrion & Ulrich Mansmann

Abstract

Background: There is a plethora of treatment strategies for Menière's disease, including endolymphatic sac decompression, restriction of salt and fluid intake, diuretics, intratympanic injections of gentamycin, administration of corticosteroids, and medical treatment with betahistine-dihydrochloride. There are, however, no state-of-the-art treatment studies in this field. The aim of this randomized placebo-controlled Phase III trial is to evaluate the effects of betahistine-dihydrochloride in a dosage of 2 × 24 mg/day, versus 3 × 48 mg/day, versus placebo on the incidence of Menière's attacks. Secondary objectives are to assess the median duration and severity of attacks as well as vestibular and audiological functions. The clinical aims of this study are to stop vertigo, reduce or abolish tinnitus, and preserve or even reverse hearing loss.

Objective: To develop and report a pre-determined statistical analysis plan (SAP) which the investigators will adhere to in analyzing the final data from the trial.

Results: BEMED is an investigator-initiated, long term, multicentre, randomized, double blind, placebo controlled, 3-arm parallel-group superiority trial that investigates and compares the effect of betahistine administered in two different dosages with placebo. Primary efficacy outcome was the number of Menière's attacks per 30 days measured by a paper-based event-driven vertigo diary over a 9-month treatment period. The original subjective patient ratings were evaluated by trained professionals of a central blinded endpoint adjudication committee according to a consensus document (SOP '*Diary Assessment*') in order to define conclusive efficacy data. A SAP for the BEMED trial was developed, which allows a comprehensive and detailed description of baseline characteristics, features of the evaluation process of the diary-based patient-reported outcome (PRO) data, different analysis sets, and the pre-determined statistical assessment of relevant outcome measures in a way that is transparent, available to the public, verifiable and preplanned before the actual analyses of trial data.

Conclusion: This detailed SAP was written prior to any analyst having access to any unblinded data and was approved by the coordinating investigator and sponsor delegated person. This document comprehensively describes the data captured by case report forms, patients' vertigo diaries and self-administered questionnaires. Its publication will ensure that confirmatory analyses are in accordance with an a priori plan related to the trial objectives and not driven by knowledge of study findings in order to minimize future analysis bias.

Trial registration: EudraCT number: 2005-000752-32; Current Controlled Trials number: ISRCTN44359668

Keywords: Menière's disease; Betahistine; vertigo attacks; vertigo diary; patient-reported outcome (PRO); statistical analysis plan (SAP); count data; randomized controlled trial

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

MEDIZINISCHE FAKULTÄT
INSTITUT FÜR MEDIZINISCHE INFORMATIONSVERARBEITUNG, BIOMETRIE UND EPIDEMIOLOGIE - IBE
LEHRSTUHL FÜR BIOMETRIE UND BIOINFORMATIK

Statistical Analysis Plan (SAP)

BEMED Trial

TRIAL FULL TITLE (Acronym)	**Medical treatment of Menière's disease with betahistine: a placebo-controlled, dose-finding study (*BEMED*)**
EudraCT Number	2005-000752-32
ISRCTN Number	ISRCTN44359668
Serial number at source	04T-617
Protocol version; Date	Protocol amendment number: Version 6; October 07, 2011
SAP Version; Date	Final Version 1.0; July 20, 2014
SAP Author	Christine Adrion, MPH
SAP Reviewer	Prof. Dr. Ulrich Mansmann

PRINCIPAL INVESTIGATOR:
Prof. Dr.med. Michael Strupp, FANA, FEAN
Ludwig-Maximilians-Universität
Dept. of Neurology, and German Center for Vertigo and Balance Disorders (DSGZ)
Klinikum Großhadern
Marchioninistr. 15
D-81377 München

TRIAL STATISTICIAN:
Prof. Dr.rer.nat. Ulrich Mansmann
IBE - Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie,
Lehrstuhl für Biometrie und Bioinformatik
Ludwig-Maximilians-Universität München
Marchioninistr. 15
D-81377 München

Approval of SAP and Signature Page

I have carefully read this statistical analysis plan, SAP Version 1.0, and agree to the described methods and proceedings.

Prof. Dr.rer.nat. Ulrich Mansmann	Responsible Biometrician, Data Management	Date
Prof. Dr.med. Michael Strupp	Coordinating Investigator & Sponsor	Date
Christine Adrion	Biometrician	Date

Statistical analysis plan for a multicenter, double-blind, randomized, placebo-controlled trial on betahistine for the treatment of Menière's disease

Adrion C

Institute for Medical Information Sciences, Biometry and Epidemiology (IBE),
Ludwig-Maximilians University, Marchioninistr. 15, 81377 Munich, Germany

Table of Contents

Abbreviations and Definitions

AE	Adverse Event
AEP	acoustic evoked potential
BDRM	Blinded data review meeting
BL	baseline
CI	confidence interval
CRF	Case Report Form
DHI	Dizziness Handicap Inventory
DM	data management
DSMB	Data Safety Monitoring Board
FAS	Full Analysis Set
FCS	fully conditional specification
FU	follow-up
GLMM	generalized linear mixed model
HD	high dosage group
HLT	Higher Level Term
IBE	Institute for Medical Information Sciences, Biometry and Epidemiology, LMU Munich
IMP	Investigational Medicinal Product
ITT	Intention-to-treat
LD	low dosage group
M(C)AR	Missing (Completely) At Random
MedDRA	Medical Dictionary for Regulatory Activities
Mini-TBF12	Mini Tinnitus questionnaire (based on 12 items of the full tinnitus questionnaire)
MNAR, NMAR	Missing Not At Random
NB	negative binomial
PL	placebo group
PP	Per-Protocol
PRO	patient-reported outcome
PT	Preferred Term
QoL	Quality-of-life
R	software package R (www.r-project.org)
SAE	Serious Adverse Event
SAF	Safety
SAP	Statistical Analysis Plan
SAS	Statistical Analysis System®
SOC	System Organ Class
SOP	Standard Operating Procedure
SUSAR	Suspected Unexpected Serious Adverse drug Reaction
T(B)F	Tinnitus-Beeinträchtigungs-Fragebogen
TEAE	Treatment Emergent Adverse Event
TI	Telephone interview
VAS	Visual Analogue Scale
VDADL	Vestibular Disorder Activities of Daily Living Score

0 Introduction

0.1 Preface

The purpose of the Statistical Analysis Plan (SAP) is to ensure the credibility of the study findings by pre-specifying the statistical approaches to the analysis of study data prior to the data base hard lock and unblinding of the BEMED trial data. To prevent outcome bias and selective reporting, a detailed SAP is presented in order to avoid post hoc decisions that may affect the interpretation of the results of the statistical analyses of final data.

This SAP is a technical extension to the clinical study protocol (Amendment Version 1.6, dated October 07, 2011) and follows the principles of the guidelines International Conference on Harmonization (ICH) E3, E6 and E9, and the relevant Standard Operating Procedures (SOPs) of the IBE, in particular SOP BI03.

0.2 Background

Menière's disease is a disorder of the inner ear membranous labyrinth characterized by paroxysmal vertiginous attacks, fluctuating sensorineural hearing loss, aural fullness, and tinnitus [1, 2]. With an incidence of 7.4% it ranks 6th in frequency of all disorders diagnosed at the German Center for Vertigo and Balance Disorders [3]. The incidence of Menière's disease in a general population has been estimated as 157 per 100,000 persons in the United Kingdom [4] with a slight female preponderance (1.3 to 1). The peak age of onset is during the fifth and sixth decade [5].

The defining symptoms of Menière's disease according to the American Academy of Otolaryngology – Head and Neck Surgery consist of two or more spontaneous episodes of rotational vertigo each lasting 20 minutes or longer, hearing loss documented by audiograms on at least one occasion and tinnitus or aural fullness in the affected ear [1]. Especially in the early phase of the disease, however, patients may display only a subset of these symptoms, vertigo being the most common one (96.2% according to Paparella *et al*. [5]), followed by tinnitus (91.1%) and ipsilateral hearing loss (87.7%). The latter typically affects low frequencies but becomes more generalized as the disease progresses. In about one third of patients, the attack is preceded by an "aura" of aural fullness, worsening tinnitus or hypacusis [2]. In the remainder, the attacks occur spontaneously, at times in unrelenting clusters. Although spontaneous remissions are observed, most patients develop one or more persistent deficits, i.e. hypacusis, tinnitus or vestibular imbalance. Patients suffering from Menière's disease have been shown to suffer serious impairments in quality of life and to have an above-average risk of developing depression and anxiety disorders [6, 7].

The underlying pathophysiology of Menière's disease is commonly seen in a hydrops of the endolymphatic space of the membranous labyrinth, resulting in recurrent ruptures of the endolymphatic sac and spillage of potassium-rich fluid into the perilymphatic space [8-11]. This change of the ionic environment leads to depolarization of the vestibular nerve, thereby causing attacks of severe vertigo. The chronic deterioration of inner ear function with progressive hypacusis and tinnitus is thought to be caused by repeated exposure of the eighth nerve to high-concentration potassium [12]. A variety of possible causative factors have been associated with the evolution of Menière's disease. Among these are hypoplasia of the endolymphatic sac [13, 14], inflammation of the endolymphatic sac [15, 16], autoantibodies [17, 18], viral infection [10, 19] and vascular pathology [20].

The therapy of Menière's disease should aim at stopping vertigo, reducing or abolishing tinnitus, and preventing or even reversing hearing loss. Traditionally, medical treatments for Menière's disease aim at decreasing production and increasing absorption of endolymph. Approaches used for this purpose include salt-restriction and diuretic agents (e.g. hydrochlorothiazide). However, although several studies report relief of vestibular symptoms in many patients undergoing diuretic therapy [21-23], few data exist to support an effect on auditory acuity or tinnitus.

In the light of a possible inflammatory aetiology of Menière's disease, anti-inflammatory agents such as corticosteroids have been used. However, few data from clinical trials exist and a recent double-blind placebo-controlled study did not show any superior effect of intratympanically injected dexamethasone over placebo [24].

Effective control of vertigo can be expected by destruction of vestibular hair cells via intratympanic injection of gentamicin [2, 25]. Although low-dose regimens have been shown to reduce the frequency of hearing loss, this invasive therapeutic approach should be considered as a last resort. The same pertains to destructive operative approaches such as vestibular neurectomy or labyrinthectomy [26].

More recently, betahistine-dihydrochloride has come to be used as an alternative medical treatment in Menière's disease. Clinical studies have demonstrated its beneficial effects on the vestibular and to a lesser degree on the audiological symptoms. All these trials feature low to moderate doses of betahistine. With clinical evidence pointing towards a role of high-dosage regimens in the treatment of Menière's disease, the BEMED trial, a prospective randomized double-blind placebo-controlled dose-defining clinical trial, was conducted.

BEMED is a pragmatic trial measuring the clinical effectiveness of up to a 9 month treatment period with betahistine, assessing whether this intervention can improve the long-term outcome measured by the frequency of attacks in patients suffering from Menière's disease.

0.3 Purpose of the analyses

The statistical analyses described in this SAP will assess the efficacy and safety of betahistine-dihydrochloride in a dosage of 24 mg 2 × day (low dosage arm) and 48 mg 3 × day (high dosage arm) in comparison with placebo, and will be included in the final clinical study report or a peer-reviewed publication.

1 Study Objectives and Endpoints

1.1 Study Objectives

(ICH E3; 8.)

1.1.1 Primary objective

The primary aim of the BEMED trial is to evaluate the effect of betahistine in a dosage of 48 mg three times per day (high-dosage arm) compared to a standard dosage of 24 mg two times per day (low dosage arm) and to placebo on the absolute number of evaluated Menière attacks during the last three months of a nine months continuous treatment period. It shall be analyzed whether there is a positive effect of betahistine on Menière's disease at all, and the appropriate dosage shall be determined.

Study hypothesis:

High-dose betahistine (3 × 48 mg per day) is more effective in reducing the number of vertigo attacks in Menière's disease than low-dose betahistine (2 × 24 mg) or placebo.

The null hypothesis H_0 is defined as follows:

There is no difference in the number of evaluated Menière attacks observed during the time period 7, 8, and 9 between the three treatment groups.

1.1.2 Secondary objectives

Secondary objectives are to evaluate the tolerance and side effects of the novel high dosage of betahistine, the effect of different dosages on duration and severity of Menière attacks, vestibular and audiological function or deficits like hearing loss and tinnitus (in the selected ear), as well as on the handicap in daily living activities due to Menière's disease.

1.2 Endpoints

(ICH E9; 2.2.2)

1.2.1 Primary efficacy endpoint

Primary efficacy endpoint is the absolute number of Menière attacks recorded by a patient vertigo diary. The *primary efficacy outcome measure* is the absolute number of Menière attacks during the last three months of a nine month treatment period, i.e. within the defined time intervals 7, 8, and 9 (i.e. between day 181 and 270). The time unit is 30 days, starting from a time point 1 defined as the date of first intake (with the day of first study drug intake being Day 1) – as considered appropriate after the BDRM (see section 6.1.1). The primary analytic objective is to quantify the attack incidence within a 90-day period at the end of the 270 day treatment period and to compare between the three treatment groups.

1.2.2 Secondary efficacy endpoints

Diary-based secondary endpoints within time interval 7, 8, 9 (i.e. between day ≥ 181 and ≤ 270 of the 9-month treatment period):

1. Median *duration* of Menière attacks during the last 3 months of the treatment period
2. Median *severity* of Menière attacks during the last 3 months of the treatment period.

Secondary endpoints measured during office visits:

The derived variable "selected ear" is defined in section 1.3.3.

3. *Peripheral vestibular function* determined by electronystagmography (ENG) under caloric irrigation (two test conditions for right and left ear: 30 °C for the cool irrigation, 44 °C for the warm irrigation): will be used as secondary endpoint. Difference between treatment groups in absolute change in the *angular velocity* for the caloric nystagmus response (recorded in °/sec) between baseline and 9-month visit for the selected ear will be assessed.
4. Absolute change of *audiometrically assessed hearing loss* between baseline and 9-month visit: For test condition 250 Hz, 500 Hz, 1000 Hz, and 4000 Hz, respectively, the decibel [dB] will be

assessed for the selected ear during bone conduction[1].

5. *Tinnitus intensity* [db] determined by audiometry (for right and left ear): As secondary endpoint the absolute change between baseline and 9-month visit will be defined for the selected ear.
6. ~~*Objective hearing loss,* determined by acoustic evoked potentials (AEPs) for right and left ear: As secondary endpoint the absolute change between baseline and 9- month visit, will be defined for Peak I, II, III~~ [Statistical analysis is not possible due to insufficient data quality, and a huge amount of missing data (examination not performed)].

The following three secondary endpoints are based on QoL patient questionnaires (dizziness and self-assessment scores)

7. Handicap/ impairment due to vertigo or tinnitus, assessed by the *Dizziness Handicap Inventory* (DHI), the *Vestibular Disorders Activities of Daily Living* (VADL), and the *Mini-TBF12 score*: Absolute change between baseline and 9-month visit.

Visit 4 and its relation to the general time axis of the trial:

For statistical analyses of secondary endpoints measured during office visits we define as visit 4 (9 months visit) the latest measurement between treatment day 240 to 300 after baseline visit (i.e. day 270 ± 30 days).

1.2.3 Safety endpoints

[Details will be presented by our partner ABBOTT.]

Safety will be evaluated with a summary of

- AEs, SAEs, SUSARs classified with the following covariates: severity, frequency, causality ("definite/certain", "probable/likely", "possible", "unlikely", "no relationship", "not assessable/unclassified"), action taken, outcome.

as well as the following

- laboratory safety parameters:
  potassium, sodium, creatinine, CRP, glutamat-oxalacetat-transaminase (GOT), glutamat-pyruvat-transaminase (GPT), gamma-glutamyl-transferase (Gamma-GT), Blood glucose level, hematocrit, hemoglobin, erythrocytes, leucocytes, thrombocytes.

The number of occurrences of any AEs, SAEs, or SUSARs, which are classified as certainly, probably, or possibly related to the treatment, will serve as safety measures. Especially the following signs and symptoms are considered to be important:

- flush
- novel/severe vertigo or dizziness
- tachycardia
- severe persisting headache

[1] The test condition "air conduction" is not suitable to define a key secondary endpoint concerning audiometrically assessed hearing loss.

- hypotonia (systolic blood pressure < 100mmHg)
- increase of alalnine aminotransferase level > two times the upper limit of the normal range or higher
- bronchospasm
- Quincke's edema (edema of the upper respiratory tract or the mucosa)

at any time of the entire study period.

1.3 Derived variables

1.3.1 Primary efficacy outcome "number of evaluated attacks"

The number of *evaluated* Menière attacks is derived from the original patient-reported outcome (PRO) data recorded in the vertigo diaries. These subjective diary-based patient ratings were evaluated by trained professionals according to a consensus document in order to define conclusive efficacy data from a clinical perspective. The SOP describing the process of diary assessment in order to derive the primary efficacy outcome is an official part of this SAP (see APPENDIX I: SOP '*Diary Assessment*').
Several complex diaries have to be evaluated on a patient-individual manner by a clinician. These patients will be addressed during the blinded data review meeting.

1.3.2 QoL: Dizziness and self-assessment questionnaires

1.3.2.1 VDADL score

To determine how well patients judged their functional compensation, they completed questionnaires designed for vestibular patients that included the vestibular disorders activities of daily living (VDADL) scale. The VDADL consists of 28 questions that assess subjects' comfort and ability to perform activities categorized as *functional (F), ambulatory (A),* and *instrumental (I)*, as well as a "total scale" that summarizes all three categories. In the original definition of the VDADL, subjects score their responses to each question using integer numbers ranging from 1 ("`best`") to 10 ("`worst`").

According to Cohen & Kimball (2000) the measured parameter to summarize the 3 subscales and the total score is the median score. As secondary outcome the total VDADL score, i.e. the median value of answers across all 28 questions will be used. Additionally, the 3 VDADL subscores are derived by determining the median of the corresponding items.

In this way, if the patient fails to answer a question (no matter if the last column ("`[NA], keine Antwort`") is ticked or not), the VDADL score is not affected significantly by missing values. Unlike the mean, the median is not unduly influenced by extreme answers that do not agree with the remainder of the subject's assessment and avoids the bias that would be introduced into a sum if a subject omits an answer or uses the non-applicable rating ("`NA`").

1.3.2.2 DHI score

To assess the impact of impairment the patients are asked to fill out the 25 item DHI questionnaire.
The original DHI total score (range: 0 to 100 points) consists of three subscales: *functional subscale (F), emotional subscale (E) and a physical subscale (P)*. The top score is 100 (maximum perceived disability), the bottom score is 0 (no perceived disability).

The subjective measure of the patient's perception of handicap due to the dizziness can be categorized as follows (Jacobson & Newman, 1990):

- 16–34 Points (mild handicap)
- 36–52 Points (moderate handicap)
- 54+ Points (severe handicap)

For each of the 25 items, a "`yes/always`" response is scored 4 points, a "`sometimes`" response 2 points, and a "`no`" response 0 points.

To deal with missing items, we use the *derived mean DHI score* (**DHI_Total$_{mean}$**) as outcome variable averaging for the number of answered questions:

$$\text{DHI\_Total}_{\text{mean}} = \left({}^{1}/_{\sum_i item_i \neq NA}\right) \sum_{i=1}^{25} item_i$$

whereas NA denotes a missing answer. In `R` code this means: `mean(., na.rm = T)`.

1.3.2.3 Mini-TBF12 score

The full tinnitus questionnaire (TQ) of Goebel and Hiller (1994) measures the impairment due to tinnitus with six partially correlating factors and is a standardized instrument for grading the severity of tinnitus.

Instead of using the full TF global score (for which 40 of the 52 items of the TF are needed for computation of the total score), the **Mini-TF12 score** according to Hiller & Goebel (2004) as an abridged and more compact measure will be analyzed to assess tinnitus-related psychological distress. The following selected 12 items reflect most central and characteristic aspects and will be used to calculate the Mini-TBF12 score[2]:

- [5] Ich bin mir der Ohrgeräusche vom Aufwachen bis zum Schlafengehen bewusst.
- [16] Ich mache mir wegen der Ohrgeräusche Sorgen, ob mit meinem Körper ernstlich etwas nicht in Ordnung ist.
- [17] Wenn die Ohrgeräusche andauern, wird mein Leben nicht mehr lebenswert sein.
- [24] Auf Grund der Ohrgeräusche bin ich mit meiner Familie und meinen Freunden gereizter.
- [28] Ich sorge mich, dass die Ohrgeräusche meine körperliche Gesundheit schädigen könnten.
- [34] Wegen der Ohrgeräusche fällt es mir schwer, mich zu entspannen.
- [35] Oft sind meine Ohrgeräusche so schlimm, dass ich sie nicht ignorieren kann.
- [36] Wegen der Ohrgeräusche brauche ich länger zum Einschlafen.
- [39] Wegen der Ohrgeräusche bin ich leichter niedergeschlagen.
- [43] Ich denke oft darüber nach, ob die Ohrgeräusche jemals weggehen werden.
- [47] Ich bin Opfer meiner Ohrgeräusche.
- [48] Die Ohrgeräusche haben meine Konzentration beeinträchtigt.

Each item can be answered as either "`true`" (= 2 points), "`partly true`" (= 1 point) or "`not true`" (= 0 points). The crude Mini-TBF12 score is the sum of all points, ranging from 0 to 24.

[2] http://www.tinnitus-liga.de/pages/sonstiges/aktionsleiste/tinnitus---test/tinnitus-testbogen.php

According to section 1.3.2.2 we use the *derived mean Mini-TBF12 score* (**MiniTF$_{mean}$**) as outcome variable averaging for the number of answered questions defined above (item number #5, 16, 17, 24, 28, 34, 35, 36, 39, 43, 47, 48) ignoring the missing values

$$\text{MiniTF}_{\text{mean}} = \left(1/\sum_i item_i \neq NA\right) \sum_{i \in \{5,16,17,24,28,34,35,36,39,43,47,48\}} item_i$$

whereas NA denotes a missing answer. In `R` code this means: `mean(., na.rm = T)`.

1.3.3 Selected ear

According to the inclusion criteria, a study participant suffers from audiometrically documented hearing loss either in the left or right ear, or both ears. Additionally, tinnitus or aural fullness in the treated ear has to be diagnosed prior to enrolment. The *selected ear* (variable `selectedear` in the dataset `ear`) chosen for statistical analyses is defined as follows:

- For patients with audiometrically documented hearing loss either in the left or right ear, the selected is the left or right ear, respectively.
- For patients with audiometrically documented hearing loss in both ears and documented tinnitus/aural fullness in either the left or right ear, the selected ear is the single left or right ear affected by tinnitus/aural fullness.
- For patients with audiometrically documented hearing loss in both ears and documented tinnitus/aural fullness in both ears, the selected ear will be chosen randomly.

For a detailed description see the `R` code in the section 10.6.

2 Study Methods

2.1 General Study Design

(ICH E3;9)

BEMED is a pragmatic trial of the clinical effectiveness of up to a 9 month treatment period with betahistine and comprises three arms:

1. placebo (PL)

the active drug with 2 different dosages:

2. therapy with low-dose (LD) betahistine (2 x 24 = 48 mg),
3. therapy with high-dose (HD) betahistine (3 x 48 = 144 mg).

Study configuration and experimental design:

- investigator-initiated, longitudinal, multicenter, double-blind, randomized, placebo-controlled, 3-arm parallel-group phase III superiority trial,
- confirmatory dose-defining study
- fixed sample design
- method of treatment assignment: block randomization with stratification by site.

2.2 Inclusion-Exclusion Criteria and General Study Population

(ICH E3;9.3. ICH E9;2.2.1)

This section is intended to describe particulars about all of the subjects in the study. It is distinct from the Analysis Population (section 4.2). This section is intended to describe the intended characteristics of ***all*** the subjects in the study.

Patients were enrolled only if they meet all of the following **inclusion criteria**:

- Diagnosis of definite Menière's disease:
  - Two or more definitive spontaneous episodes of vertigo of 20 minutes duration or longer
  - Audiometrically documented hearing loss on at least one occasion
  - Tinnitus or aural fullness in the treated ear
  - Other causes excluded
- At least two attacks per months for at least three subsequent months
- Age 18 to 80 years
- Written informed consent signed and dated by the patient (or patient's authorized representative) and by the person obtaining the consent, indicating agreement to comply with all protocol-specified procedures.
- Female patients of childbearing potential must have a negative pregnancy test within 7 days before initiation of therapy. Postmenopausal woman must be amenorrheic for at least twelve months

Exclusion criteria:

General criteria

- Participation in another study with an investigational drug or device within the last 30 days, prior participation in the present study or planned participation in another trial
- Women known to be pregnant or lactating
- Woman of childbearing potential who are not willing to practice acceptable methods of birth control (during and for three months after therapy) to prevent pregnancy.

Concerning vertigo/ dizziness

- Other vestibular disorder such as
  - vestibular migraine
  - phobic postural vertigo
  - benign paroxysmal positioning vertigo
  - paroxysmal brainstem attacks

- Contraindications for the treatment with betahistine, such as
  - bronchial asthma
  - pheochromocytoma
  - pregnancy or breast-feeding
  - severe dysfunction of liver or kidney
  - ulcer of the stomach or duodenum
  - treatment with other antihistaminic drugs

Safety related criteria

- severe coronary heart disease or heart failure
- persistent hypertension with systolic blood pressure > 180 mmHg or diastolic BP > 110 mmHg (mean of 3 consecutive arm cuff readings over 20-30 minutes) that cannot be controlled by antihypertensive therapy

Potentially interfering with outcome assessment

- life expectancy < 12 months
- other serious illness, e.g. severe hepatic, cardiac or renal failure, acute myocardial infarction, neoplasm or a complex disease that may confound treatment assessment

Co-medication

- treatment with other antihistaminic drugs

2.3 Randomisation and Blinding Methodology

(ICH E3; 9.4.3, 9.4.6. ICH E9; 2.3.1, 2.3.2)

Blinding

Betahistine was encapsulated using mannitol and aerosile as filling material. The modification was performed by the Pharmacy of the University Hospital Heidelberg. Betahistine was refilled from original pharmacy packaging to vials under sterile conditions and relabeled.

To ensure similarity of interventions, the placebo drug was matched to the study drug for taste, color, and size. To be more detailed, placebo was an identically appearing capsule filled with mannitol and aerosil according to DA. Placebos were also refilled to vials.

Randomization procedure

The concealed 1:1:1 allocation was an internet-based randomization schedule (https://wwwapp.ibe.med.uni-muenchen.de/randoulette) stratified by site. Details concerning block size will be provided in the final study report and the main publication.

2.4 Study Variables and Study Schema

(ICH E3; 9.5.1. ICH E9; 2.2.2)

Details concerning the treatment and follow-up (post treatment) period together with the frequency and timing of relevant variables or assessments are displayed in Table 1.

Table 1. Schedule of enrolment, interventions and assessments. BL, V1, V2, V3, V4, FU: office visits. T1, T2, T3, T4, T5: telephone visits. BL = Baseline, FU = Follow-up.
Dizziness or tinnitus self-assessment-scales: VDADL, DHI, tinnitus questionnaires.

		Treatment Period									
	Baseline (Day 1) BL (V0)	**Month 1:** V1	**Month 2:** T1	**Month 3:** T2	**Month 4:** V2	**Month 5:** T3	**Month 6:** V3	**Month 7:** T4	**Month 8:** T5	**Month 9:** V4	**Month 12:** FU
Informed consent	×										
Eligibility screen	×										
Randomisation	×										
Medical history	×										
Vertigo diary		×	×	×	×	×	×	×	×	×	×
Dizziness/Tinnitus Self-assessment scales: DHI, VDADL, Mini-TBF12	×	×			×		×			×	×
Physical / neurological examination	×	×			×		×			×	×
Blood sample	×	×			×		×			×	×
Electronystagmography (ENG)	×						×			×	×
Neuro-orthoptic examination	×						×			×	×
Acoustic evoked potentials (AEPs)	×						×			×	×
Audiometry, Tinnitus intensity	×						×			×	×
Delivery of trial medication	×	×			×						
Treatment compliance, drug counting		×	×	×	×	×	×	×	×	×	×
Concomitant medication	×	×	×	×	×	×	×	×	×	×	×
(S)AE monitoring		×	×	×	×	×	×	×	×	×	×

For the primary efficacy analysis, non-scheduled office or telephone visits are not an issue, since 30 day intervals will be defined for attack data.

3 Sample Size and Sample Size Re-estimation

(ICH E3; 9.7.2. ICH E9; 3.5)

A total of 14 study sites participated in the recruiting process.

Initially planned sample size

The sample size calculation was based on the Wilcoxon (Mann-Whitney) rank-sum test. Therefore, three parameters are relevant: the level of significance, the power of the two-sided test and the probability that an observation in Group A is less than an observation in Group B. Based on pilot data (27 patients),

the probability that an observation in Group A is less than an observation in Group B was estimated to be 0.9 with a 95% confidence interval of [0.75; 0.98]. If the sample size calculation is based on the lower bound of the 95% confidence interval for the parameter of interest, a sample size of 21 in each group will have 80% power to detect the difference between both groups using a Wilcoxon (Mann-Whitney) rank-sum test with a 0.05 two-sided significance level (Software used: nQueryAdvisor Version 6.0).

On the basis of our experience with patient compliance in previous studies and routine treatment, we observed a drop-out rate of about 45% to 50%. This study will implement a close contact between study investigator and patient which motivates the patient to stay within the study. Therefore, we believe to be able to reduce the drop-out rate below 20%. Thus, a total of 84 patients (28 in each treatment group) have to be enrolled.

Revised sample size calculation

Due to uncertainty about the dropout rate and after finding a lower pre-randomization baseline rate for the attack frequency using data from 19 study patients allocated to the BEMED trial (mean baseline attack frequency was 7 attacks), a *blinded sample size re-calculation* was performed.

Primary efficacy endpoint is the number of Menière attacks in the three treatment arms during the last 3 months of the 9 month treatment period. This outcome variable is skewed and therefore cannot be considered to be normally distributed.

An overall effect of treatment is analyzed with a longitudinal approach based on a linear random intercept model for the arcus-sinus-hyperbolicus transformed frequency measurements. Recalculation used data from an open, non-masked trial published in Strupp *et al*. (2008) (112 patients), and, additionally, baseline data for the primary outcome measured for study patients allocated to the BEMED trial (19 patients). Based on these two data sources, the mixed modelling approach identified a time effect of -0.06 and an effect of medication on the number of attacks in the course of time of about -0.08 (transformed scale). The individual variation of baseline level (i.e. standard deviation of predicted random intercepts) was estimated to be 0.8, the within-error to be 0.5.

Using the combination between model and observed baseline variation it was possible to determine the new planning figures for a sample size re-estimation by simulation:
With the parameter estimates from the mixed modelling approach and the mean baseline attack frequency on the transformed scale, data for number of attacks could be derived for month 0 and 12 for both treatment groups A and B (sample size for both groups A and B was 1000). The protocol performs the sample size calculation for a Mann-Whitney U-test between the differences of baseline and final attack frequency after 12 months in treatment groups (Δ_A, Δ_B). Based on the simulation scenario described above it was possible to determine the relevant parameter, $P[\Delta_A > \Delta_B]$, as 0.33.

Based on this parameter, a sample size of 46 in each group (i.e. a total of 138 patients in the whole study) will have 80% power to detect the difference between both groups using a Wilcoxon (Mann-Whitney) rank-sum test for two independent groups with a 0.05 two-sided significance level (Software used: nQuery Advisor Version 7.0).

On the basis of pilot data on patient compliance and due to the fact that this study will implement a close contact between study investigator and patient which motivates the patient to stay within the study, we assumed a drop-out rate of approximately 25%. Hence, a total of 186 (62 in each treatment group) had to be enrolled to the trial. It has to be taken into consideration that about 50% of patients fulfilling the inclusion criteria for this trial might refuse to give their consent to participate in this trial,

because the frequency of study visits is high and the medication might consist of placebo for an entire 9 months. We therefore expected to screen about 372 patients for eligibility.

4 General Considerations

4.1 Timing of Analyses

Participant recruitment was completed in November 2012, and final participant follow-up was completed in November 5, 2013. The end of the trial is defined as the date of the last visit of the last patient undergoing the trial (LPLV: 05.11.2013).

All final analyses will be performed on the *derived database.* After having documented all CRF data and after data cleaning and query resolution have been completed, the following prerequisites for unblinding have to be fulfilled:

- blinded data review
- resolution of all queries concerning CRF, diaries and questionnaires
- the finalization and approval of this SAP document.

All these processes before data base locking must take place to comply with requirements documented in the IBE-SOP DM07, DM08 and DM11.

The statistical analysis plan was completed and signed as approved by the study investigators in July 2014. Following data integrity checks the database will be locked end of July 2014 and the statistical analyses specified in the SAP will be performed in August 2014.

4.2 Analysis Populations

(ICH E3; 9.7.1, 11.4.2.5. ICH E9; 5.2)

This section is designed to identify the characteristics needed for inclusion in particular populations used in the analyses.

During the BDRM the exact process for assigning each subject's inclusion or exclusion status will be defined and documented prior to breaking the blind along with any predefined reasons for eliminating a subject from a particular population.

4.2.1 Full Analysis Set (FAS)

The primary efficacy analysis follows the principle of intention to treat (ITT), which implies that study data are analyzed based on the original allocation of study participants, regardless of a treatment received.
Withdrawals, participants lost to follow-up and participants who did not adhere fully to the study protocol will not be excluded from the primary efficacy analyses provided that they satisfy major entry criteria). Also, a patient will contribute to the primary efficacy analysis provided that she/he contributed attack data. This is defined between the BDRM and database closure (see APPENDIX section 12).

Explicit statements about post-randomization exclusions:

Hence, the full analysis set (FAS) population includes all subjects randomized (irrespective whether they were treated or not), and who do not fail to satisfy a major entry criteria[3]. This assessment is part of the minutes of the blinded data review meeting.

The exclusion of subjects who failed to satisfy one or more major entry criteria is justified because the entry criteria were measured prior to randomization. The exclusion of subjects who took no study medication is justified because the decision of whether or not to begin treatment could not be influenced by knowledge of the assigned treatment. The exclusion of subjects without primary **and** secondary efficacy data is required because of the models that will be applied and requires the assumption of missingness at random.

4.2.2 Per Protocol Set (PP)

The PP set consists of all subjects who did not substantially deviate from the protocol as to be determined on a per-subject basis at the BDRM before final data base lock. The PP set of subjects defining a subset of the FAS is characterized as follows (also see section 5.2):

All subjects from the FAS

- for whom no major protocol violations were detected (e.g. poor compliance, errors in treatment assignment, etc.). This assessment is part of the BDRM.

AND

- who are under **treatment at least 8 months, i.e. ≥ 240 days**, counting from day of first intake [*completion of a certain pre-specified minimal exposure to the treatment regimen*]

AND

- who provide **diary information** within the primary time intervals {7, 8, 9} after the defined starting point, regardless of the number of evaluated days within the 30-day time intervals {7, 8, or 9} [*availability of measurements of the primary variable within the time period of interest*].

Hence, patients who prematurely discontinue from the study or treatment before time interval 7 will be excluded from the PP analysis set.

4.2.3 Safety Population (SAF)

- All subjects who received any study treatment (including control) and for whom post-start-of-study-treatment safety data are available, but excluding subjects who drop out prior to receiving any treatment. Therefore, subjects who are confirmed as providing follow-up regarding adverse event information are part of the SAF.

The safety population is not equivalent to the FAS population. The safety population also includes patients receiving any treatment but not providing any efficacy data (i.e., those patients who were excluded from the primary efficacy population). The FAS population is a therefore a subset of the Safety population.

[3] e.g. attack history before study enrolment

4.3 Covariates and Subgroups

(ICH E3; 9.7.1, 11.4.2.1. ICH E9; 5.7)

There exists no a priori hypothesis of subgroup differences. Hence, no pre-planned confirmatory subgroup analyses will be performed to explore evidence for a difference in treatment effects (interaction effect).

Exploratory subgroup-specific *summary* statistics will be reported for gender and age:

- gender
- age, coarsening into ≤45, (45, 55], (55, 65], >65. Further cut-offs will be investigated post-hoc.

Subgroup analysis will be performed by studying the interaction effect between treatment and covariates.

4.4 Missing Data

(ICH E3; 9.7.1, 11.4.2.2. ICH E9;5.3. EMA Guideline on Missing Data in Confirmatory Clinical Trials, 2010; NRC Report 2010)

The primary analysis is a *mixed effects modeling approach* that assumes that missingness is at random (**MAR**) for both permanent (i.e. dropout) and intermittent missing data pattern. That is, the mixed model assumes that, given the statistical model (i.e. conditional upon the independent variables in the analysis) and given the observed values of the dependent variable (i.e. the primary endpoint 'number of evaluated attacks'), the probability of missingness does not depend on the unobserved outcomes of the dependent variable.

The main model under MAR is based on the assumption that no post-randomisation variable will be predictive of the partially observed outcome. No multiple imputation techniques will be performed for primary efficacy analysis which is based on an "all observed data approach", and therefore *is* optimally statistically efficient.

Specific missing data like *date for end of treatment* will be handled as follows:

If the exact date is not known, but month and year is reported, the exact date will be defined as `15.mm.yyyy`, i.e. day=15.
If the exact date is totally missing the date for end of therapy will be manually "imputed" (in a *non*-statistical sense). This means that the missing date will be filled in by a reasonable date (new derived variable: `stop_dat`) before the last patient contact by applying a SAS "algorithm". For more details see the SAS program in the APPENDIX, section 10.7. E.g., for early dropouts who discontinue before V1, the treatment end is set to `stop_dat= Einnahm1_dat + 1`. For dropouts after V1 the date for end of treatment will be set to the midpoint between two visits.
A few cases will need a specific consideration based on the patient's treatment history (e.g. if data are available off-treatment). For these cases, the variable `DECISION` indicates whether the imputed date `stop_dat_Nauta` or the original date `therend_dat_IBE` as documented in the SAS database must be used to derive, e.g., treatment duration (derived variable: `study_drug_duration_final`).

The variable `stop_dat_final` includes the final (imputed or original) dates for treatment end after the decision process and should be used for further analyses.

4.5 Multi-center Studies

(ICH E3;9.7.1, 11.4.2.4. ICH E9; 3.2)

The primary efficacy analysis of the multi-center BEMED trial will be performed without adjusting for center effects although center was used in the treatment allocation process. Center as main effect will be studied as one of the sensitivity analyses. The interaction between center and treatment will not be considered.
Center will not be adjusted for in the primary analysis since the BEMED trial has not been explicitly designed with enough power to detect center effects. For a relevant amount of sites the number of patients per center is too small to allow the inclusion of center as a covariate in the main model, and would introduce too many categories (this is particularly an issue for non-normal response data).

4.6 Multiple Testing

(ICH E3; 9.7.1, 11.4.2.5. ICH E9; 2.2.5)

Adjustment for multiplicity is considered necessary since the trial has a single pre-specified primary outcome measure, but 3 treatment arms. Therefore, a formal **closed-testing procedure** was adopted that examines the 3 hypotheses (H_{01}: HD vs. LD, H_{02}: HD vs. PL, H_{03}: LD vs. PL) in such a way that preserves the overall $\alpha = 5\%$ significance level of the confirmatory efficacy analyses. The closed-testing procedure consists of an overall global test testing if there is any effect at all ($H_{0,\,global} = H_{01} \cap H_{02} \cap H_{03}$) followed by the pairwise comparisons given by H_{01}, H_{02}, H_{03} using the same significance level of $\alpha = 5\%$. If the global test for $H_{0,\,global}$ will not be significant no pairwise comparisons will be valid.

The secondary outcomes are exploratory and the results will only be interpreted as supportive evidence related to the primary outcome.

5 Summary of Study Data

All continuous variables will be summarized by treatment group using the following descriptive statistics:
N (non-missing sample size), mean ± standard deviation (SD), median, maximum and minimum. The absolute frequency and percentages (based on the non-missing sample size) of observed levels will be reported for all categorical variables.

In general, *patient listings* will be sorted by subject within study center, and treatment group (Placebo, experimental low dose, experimental high dose), and when appropriate by visit number within subject.

All *summary tables* will be structured with a column for each treatment and overall in the order

- Placebo,
- Experimental Low Dose (LD)
- Experimental High Dose (HD)
- All subjects (only for baseline observations)

and will be annotated with the total population size relevant to that table/treatment, including any missing observations - unless specified otherwise.

5.1 Subject Disposition

A CONSORT diagram (CONsolidated Standards of Reporting Trials[4]) according to the CONSORT 2010 Statement will be reported to establish how many subjects reached the various stages of the trial, how many dropped out and for what reasons (death, AEs, treatment failure, withdrew consent, loss-to follow-up). For example, the number screened for eligibility, randomized, completed office visits 1, 2, 3, 4 (using dates for the physical or neurological examination), and reached study termination defined by follow-up visit V5 will be described.

5.2 Protocol Deviations

Patients with a major deviation defined below are excluded from the PP analysis set defined in 4.2.2. During the blinded data review meeting several protocol deviations will be defined and discussed.

major deviations:

- study dropout before time interval 7, and no off-treatment data provided
- **treatment duration less than 8 months**, i.e. treatment duration **< 240 days** (day 1 defined as date of first intake, see variable: `Einnahm1_dat`).[5]

5.3 Evaluation of Demographic and Baseline Variables

The following pre-treatment patient characteristics and baseline covariates will be displayed descriptively.

demographics:

- sex
- age (cut-offs: ≤45, (45, 55], (55, 65], >65)
- ethnic group
- body weight
- body height
- Body Mass Index

baseline variables recorded before randomisation or first treatment administration:

- medical history
- physical examination at baseline visit
- neurological examination at baseline visit
- laboratory parameters
- electronystagmography (ENG):

[4] http://www.consort-statement.org/

[5] The defined time interval #9 starts with Day 241.

  - spontaneous nystagmus,
  - postrotatatory nystagmus,
  - bithermal caloric test

- neuro-ophthalmologic examination
- audiometric testing: 1.) air conduction, 2.) bone conduction), tinnitus intensity
- speech audiometry: hearing loss concerning numbers (determined by formal audiometric testing), monosyllabic tests

The summary statistics will be displayed overall and stratified by treatment group. If considered appropriate these data are summarised by center as well.

5.4 Concurrent Illnesses and Medical Conditions

Medical history and adverse events will be coded using MedDRA. The summary statistics for physical and neurological examinations (pathological findings since last office visit) will be produced following the introductory part of section 5.

5.5 Prior and Concurrent Medications

Prior and concurrent medications will be coded using the WHO Drug Dictionary. The summary statistics will be reported in accordance with the introductory part of section 5.

Betahistine therapy after withdrawal or during the follow-up period will be investigated in tables and listings.

For betahistine, registered trade names are:

- Aequamen®/ Aequamen®-forte Tabletten
- Betahistin-ratiopharm® 6 mg/ 12 mg Tabletten
- Betavert® 6 mg/ 12 mg Tabletten
- Betavert® N 8 mg/ 16 mg/ 24 mg Tabletten
- Vasomotal® 16 mg/ 24 mg
- Vasomotal® Tropfen 8 mg/ml

5.6 Evaluation of Treatment Compliance and Exposure

Assessment of treatment compliance included: remaining pill count in bottle 1 and 2 documented on the CRF and diary records of medication.

Method for calculating a measure of treatment compliance

- **Treatment duration** (difference between date for end of treatment and date of first intake)

Treatment compliance will not be calculated, e.g. based on drug accountability, due to insufficient data quality and due to a high proportion of missings. Additionally, the number of capsules delivered at V1 (month 1) and V2 (month 4) was not recorded on the CRF. The number of capsules delivered at BL is not equal across patients and sites.

6 Efficacy Analyses

Unless stated otherwise, all null hypotheses will be tested at the nominal 2-sided 5% significance level. Frequentist techniques will be applied for all analyses. Since the likelihood-based approach is very complex and needs monitoring of performance and convergence issues, we choose two alternative algorithmic approaches to check the reliability of the numerical results. To this end, a Bayesian approach can be considered, and will be applied for corresponding sensitivity analyses and as a powerful tool for model validation. For further technical details see APPENDIX section 10.5 for WinBUGS code (based on MCMC sampling) and INLA code (approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations (INLAs)).

6.1 Primary Efficacy Analysis (ITT)

A MAR-based analysis excluding patients who provide no post-baseline outcome data will be performed according to an "all observed data strategy"[6]. Due to study dropouts not all patients randomized will be considered. The primary analysis population is the FAS population consisting of patients with at least one post-baseline measurement of the primary efficacy variable "absolute number of attacks" based on a certain amount of evaluable days.

We aim to address the *de jure hypothesis*, assessing the *on-treatment efficacy* (NRC Report 2010), i.e. the question is whether the treatment is effective under the best case scenario. It is to estimate the difference in outcome improvement in all randomized patients at the planned endpoint of the trial attributable to the initially randomized medication. Hence, we seek to measure the de jure estimand of treatment effect.

Efficacy data after withdrawal of randomized study medication (**off-treatment data**) will be included in the primary analysis. As sensitivity analysis, attack information will be censored[7] for patients providing primary outcomes after treatment dropout following the proposal of, e.g., Mallinckrodt, Roger *et al*. (2014), Mallinckrodt, Lin *et al*. (2012) or Keene (2011).

6.1.1 Fitting the main model

The main analysis is done unadjusted for baseline covariates or site.

The time axis is divided into *equidistant* time intervals of length 30 days ("time window") defining as starting point 1 the date of first intake.

To describe the analysis of the Menière attack frequencies, we denote the subject by i ($i = 1, \dots, 221$), and time by t ($t = \mathbf{1}, 2, 3, \dots, n_i$; best-case scenario for the primary analysis is $n_i = 9$). Hence, t is a numerical variable, and not the actual observation times for telephone or office visits are considered to define the time intervals. The vector t has no subscripts as for every single patient each time interval exactly corresponds to 30 days.

[6] White IR, Carpenter J, Horton NJ. Including all individuals is not enough: lessons for intention-to-treat analysis. *Clin Trials* 2012; 9(4):396-407.

[7] optional as further sensitivity analysis: censoring of off-treatment data collected more than two weeks after withdrawal (assuming that two weeks is approximately the wash-out period of betahistine).

The number of attacks (incidence counts) of patient i within time interval t will be denoted by $y_i(t)$. $d_i(t)$ is defined as the *number of evaluated days* for time interval t and will be used as offset variable for regression modeling. $d_i(t)$ can be interpreted as some measure of the *exposure* ("observation window") within a certain time interval. Therefore, the observation window is allowed to vary for each unit t of patient i.

Own research on model evaluation for longitudinal counts revealed that the negative binomial assumption is the distribution of choice for these data (Adrion & Mansmann, 2012).

A **negative binomial loglinear mixed model** (NB GLMM) with random intercept and random slope associated with time, and offset for the log-transformed number of evaluated days, will be applied. The linear component describes the structure of $g(\mu_i(t))$, where g is the log-link function, and $\mu_i(t)$ denotes the expected number of attacks within time interval t. The *incidence rate* $\mu_i(t)/d_i(t)$ would be the number of evaluated attacks per unit time.

For the main analysis, the linear predictor is defined as

$$\eta_i(t) = (\beta_0 + b_{0i}) + (\beta_1 + b_{1i}) \cdot t + \beta_2^{LD} I_i^{LD} + \beta_2^{HD} I_i^{HD} + \gamma_2^{LD}\left(I_i^{LD} \cdot t\right) + \gamma_2^{HD}\left(I_i^{HD} \cdot t\right) + \log(d_i(t))$$

mean structure:
The three treatment groups (PL, LD, and HD) will be dummy-coded, whereas PL will be used as reference category. I_i^{LD} and I_i^{HD} are treatment indicator variables having the value 1 in the case that patient i was randomized to treatment arm LD or HD, respectively, and 0 otherwise. Hence, the population-level parameter vector (fixed effects) consists of the intercept (β_0), the main effect for time (β_1), the main effect for treatment group ($\beta_2^{LD}, \beta_2^{HD}$), and the <u>two parameters of interest</u> (γ_2^{LD} and γ_2^{HD}), reflecting the interaction between treatment condition and time (linear). The coefficients β_2^{LD} and β_2^{HD} should be estimated rather 0 because the treatment effect is expected to happen slowly with time.
For example, the coefficient γ_2^{LD} (γ_2^{HD}) expresses the difference between the slopes of the logarithm of the average attack rate for a patient randomized to the LD (HD) group as compared to the PL group.

random effects:
$\boldsymbol{b}_i = (b_{0i}, b_{1i})'$ are random intercepts and random slopes to account for the variation among subjects both in the 'level' of Menière attacks (severity of disease) as well as in the slopes over time (i.e. random variation in slopes through time across groups specified by patient), which is assumed to be the main source of correlation among the repeated measures on the study patients.
$\boldsymbol{b}_i$ is assumed to follow a bivariate normal distribution with mean zero and an unknown precision matrix.

Underlying key assumptions:

- **MAR.** E.g., dropout due to previous recorded lack of efficacy assuming MAR means that in some sense is predictable from the observed attack data
- *linear time trend*
- *time period of primary interest* ("assessment period") for the primary endpoint is interval 7, 8, 9
- no treatment group-by-site interaction
- *no zero-inflation and no zero's hurdle* (threshold) assumed. It is <u>not</u> assumed that there is an excess of zeros, and that are two processes at work, one determining whether there are zero events or any events and the other determining the count process (ZIP, ZINB; hurdle models via truncated Poisson/ NB).

- *overdispersion parameter*: assumed to be constant and homogeneous across all 3 treatment groups, i.e. no group-specific overdispersion parameters used for model specification
- *covariance structure:* An unstructured covariance pattern will be used to model the within-patient errors
- *variance components*: Random effects precision not depending on treatment group; correlated random effects

Covariance structure

- Random intercept for patient, random slope for time.
- Random effects (b_{0i}, b_{1i}) are assumed to be normally distributed and correlated.
- we did not entertain the possibility that the person-level intercepts b_{0i} and slopes b_{1i} depend on the treatment group

Strategies to improve convergence or to avoid convergence limitations / Convergence issues:

The main model will be analyzed using a restricted maximum likelihood (REML)-based repeated measures approach. In case the NB GLMM defined above fails to converge, a heterogeneous Toeplitz covariance pattern or alternative more parsimonious correlation structures will be used in place of an unstructured one. Generally, the *first-to-converge approach*[8] (in the sense that the first structure in the ever-more parsimonious set to yield convergence) will be applied to avoid model building and hypothesis testing from the same data.

If the NB GLMM fails to converge or in case of computational difficulties (e.g. numerical instabilities), we proceed as follows: We extend the classical Poisson GLMM model which does not seem appropriate due to strict assumptions to include a per-observation error term (individual-level random intercept b_{3it}), which captures overdispersion. This type of model is often called an *"overdispersed Poisson model"* or *Poisson-lognormal model*, which is functionally similar to a negative binomial model.

6.1.2 Estimated difference in incidence of attacks within interval 7, 8, 9

The comparison of interest is the number of evaluated attacks within interval 7, 8, and 9, i.e. between day 181 and 270. The primary efficacy outcome measure is defined as the estimated mean difference between the three treatment groups *G* = PL, LD and HD in the average *incidence rate of attacks* $\lambda_i^G(t)$ *across time intervals* $t = 7, 8,$ *and* 9.

The corresponding standard errors, including the 95% confidence intervals (CIs) will be computed with a *parametric bootstrap* approach, where "parametric" means that data are simulated according to model assumptions using the estimated parameter values. At the end, bootstrap-based 95% confidence intervals are presented for *difference in (monthly) incidence rates* between HD vs. PL, LD vs. PL, and HD vs. LD.

[8] Mallinckrodt C. *Preventing and treating missing data in longitudinal clinical trials*. Cambridge 2013, p. 132

6.2 Per-Protocol analyses

All analysis will be repeated using the PP analysis set.

6.3 Secondary Efficacy Analyses

Secondary efficacy analyses (i.e. analyses concerning secondary efficacy endpoints) will be performed for the FAS as well as the PP set. The secondary efficacy outcome measures are based on the time period of primary interest.

For diary-based endpoints, the median duration or severity of attacks within interval {7, 8, 9}, i.e. between day 181 and 270, will be calculated each patient. Hence, only patients with a total number of evaluated days >0 within interval {7, 8, 9} are considered.

For secondary efficacy endpoints based on the office visit V4 the time window defined in section 1.2.2 will be applied.

6.3.1 Diary-based secondary endpoints within time interval 7, 8, 9

6.3.1.1 Attack duration

According to the SOP "Dairy Assessment" the variable duration is necessary and sufficient for a Menière attack to be assessed. Hence, there are no missing values concerning the duration of an evaluated attack being defined due to the SOP. A further prerequisite is that the duration of a patient-reported vertigo episode has to be coded with "2", "3" ,"4" or "5" in order to evaluated (i.e. attacks with duration = "1" were ignored).

	Proportion of median duration of attacks (coded)			
Treatment group	**2**	**3**	**4**	**5**
PL	n_{11}	n_{12}	n_{13}	n_{14}
LD	n_{21}	n_{22}	n_{23}	n_{24}
HD	n_{31}	n_{32}	n_{33}	n_{34}

In a first step the Jonckheere-Terpstra test as described in (StatXact for Windows, User Manual 1996) will be used to reject the global Null-Hypothesis that all three ordered treatment groups show an equal response on treatment as regards median attack duration. The Jonckheere-Terpstra test is developed to handle the situation in which a variable is measured for individuals in ordered groups and a non-parametric test for comparison across these groups is desired. If the global Null hypothesis is rejected on the significance level alpha = 5%, it is possible to perform three pair wise comparisons between the three groups again on the significance level of alpha = 5% by use of Wilcoxon Mann-Whitney U-tests as described in StatXact for Windows, User Manual (1996).

In order to quantitatively describe treatment effects together with 95% CIs we also apply a cumulative logit model. The treatment effect measure associated with this approach is the odds ratio of duration. This is the ratio of the odds of a patient treated with betahistine (LD or HD) improving to the odds of a

patient treated with PL improving. In the proportional odds model it is assumed that the cumulative odds ratio is constant across the categories of the scale used.

6.3.1.2 Attack severity

The secondary efficacy outcome measure is based on the time period of primary interest.
Only patients with a total number of evaluated days >0 within interval {7, 8, 9} are considered.
For each patient, the median severity of attacks within interval {7, 8, 9}, i.e. day 181 and 270 will be calculated.

Attack severity will be analyzed in exactly the same way as attack duration.

6.3.2 Secondary endpoints measured during office visits – (Baseline, month 9)

The difference between the three study groups in absolute change between baseline and 9-month visit, will be analyzed in a descriptive manner.

Comparison of the treatment groups for all secondary endpoints will be performed applying a t-test or Mann-Whitney U-test for quantitative measures, a Chi-square test for frequencies. Continuous variables are expressed as means ± SD if normally distributed, overall and stratified by treatment group; otherwise as median and IQR.

If the 9-month visit V4 is missing or in the case of missing baseline values multiple imputation (MI) techniques based on chained equations (MICE method[9]) assuming MAR will be applied within an ANCOVA.

6.3.3 Patient QoL questionnaires (DHI, TF, VDADL) – (Baseline, month 9)

The primary comparison is the absolute change between baseline and 9 month visit. As described in section 1.3.2 the mean scores will be used as derived variables for DHI and MiniTF score.

Hence, the mean differences

$$\Delta_9\,\text{DHI\_total}_{mean} = \text{DHI\_total}_{mean}(BL) - \text{DHI\_total}_{mean}(V4),$$

$$\Delta_9\,\text{MiniTF}_{mean} = \text{MiniTF}_{mean}(BL) - \text{MiniTF}_{mean}(V4),$$

$$\Delta_9\,\text{VDADL} = \text{VDADL}(BL) - \text{VDADL}(V4)$$

will be will be summarized with descriptive statistics overall and for each treatment group, separately. Differences between groups will be analyzed with an ANCOVA for change scores, with factor for treatment group and the baseline score as covariate, by using a closed testing approach to avoid the adjustment of the significance level because of multiple testing.

In the case of missing baseline or post-treatment values, a multiple imputation (MI) approach based on chained equations (MICE method[10]) assuming MAR will be applied. Graphical diagnostic checks on the imputed outcome values are used to check the plausibility of the imputations.

[9] Multivariate Imputation by Chained Equations

6.4 Sensitivity analyses

Sensitivity analyses aim to investigate departures from testable and untestable assumptions underlying confirmatory analyses in order assess the validity of the chosen approach. Pre-planned sensitivity analyses will be performed for the primary efficacy analysis only.

6.4.1 Model-based sensitivity analyses under MAR

6.4.1.1 Exploring testable assumptions, model checking

For the GLMM used in the primary efficacy analysis the following sensitivity analyses should be performed to check the adequacy of testable assumptions, and the impact on the estimators of interest (fixed effects of the main model):

- linear time trend
- model diagnosis (residual diagnostics, influence diagnostics)
- overdispersion
- zero-inflation
- correlation between random intercepts and random slopes

6.4.1.2 GLM for time interval {7, 8, 9}

As further sensitivity analyses assuming MAR, we only use patients with a total number of evaluated days larger than 0 across time intervals {7, 8, 9}.

This particular analysis assesses if patients who withdraw before time interval 7 show comparable efficacy results to the overall primary analysis. Marked differences would indicate strong selection processes and informative missingness.

In this analysis patients who withdraw before time interval 7 are excluded. The GLM is based on an *aggregated version* of the longitudinal approach used for the main model by summarizing the number of attacks and the number of evaluated days within time intervals 7, 8, and 9 only.
The linear predictor for the generalized linear model (GLM) will be defined according to the longitudinal model chosen for the primary analysis, leaving out the random effects part. Hence,

$$\eta_i^{7,8,9} = \beta_0 + \beta_2^{LD}\, I_i^{LD} + \beta_2^{HD}\, I_i^{HD} + \log\left(d_i^{7,8,9}\right),$$

whereas the offset $\log\left(d_i^{7,8,9}\right)$ is defined as the log-transformed absolute number of evaluated days within time intervals {7,8,9}. The linear component describes the structure of $g(\mu_i)$, where g is the log-link function according to the main model (section 6.1.1), and μ_i denotes the expected total number of attacks within time intervals {7,8,9}. The incidence rate μ_i/d_i would be the number of attacks per unit time.

The estimates resulting from this GLM approach will be compared with the estimated incidence rate derived from the GLMM approach performed for the primary efficacy analysis.

[10] The R package "mice" (Multivariate Imputation by Chained Equations in R) will be used.

6.4.1.3 Exploratory and graphical tools

To explore the impact of departures from MAR in an exploratory manner, the BLUPs (best linear unbiased predictions, i.e. the predicted random effects vector $\boldsymbol{b}_i$) resulting from the primary efficacy analysis will be displayed graphically – stratified by treatment arm – in order to investigate structural dependencies concerning the patient's dropouts status.

6.5 Exploratory Efficacy Analyses

This section includes additional analyses used for hypothesis generation and exploration and describes methods for additional analyses, such as subgroup and adjusted analyses. These analyses can also be interpreted as sensitivity analyses.

6.5.1 Adjusting for center effects

For the primary efficacy analysis the center effect was omitted assuming that it might introduce too many categories without pooling of sites (see section 4.5).

Procedures for combining of small individual sites:

As adjusted analysis the center differences with respect to response will be investigated in an exploratory manner. In the first instance, a pooling of small investigator sites <15 randomized patients will be performed to construct more usable pseudo-centers with a greater number of patients. The strategy that pools sites will also be based (1.) on the requirements for count response data to ensure convergence during the modelling stage, and (2.) on the dropout rate to ensure a reasonable number of completers per stand-alone site (O'Kelly, 2014, pp. 148, 172).
The primary model will be extended by including further fixed effects for (pooled) dummy-coded (pseudo-)centers (see section 6.1.1).
Additionally, if no convergence issues occur, interaction terms between center and treatment groups (and center and time) will be included in either the main model (GLMM) or the GLM.

6.5.2 Subgroup analyses

According to the main efficacy analysis, subgroup effects and interactions between treatment group and the baseline covariates *gender* and *age* will be explored. These exploratory subgroup analyses focus on the evidence for a difference in treatment effects, i.e. investigate for potential interaction effects.

In a first step, the main model will be extended and a main fixed effect for gender and age, respectively, will be included. Concerning age, the categorical variable according to section 4.3 will be used. In a second step, the linear predictor should be extended by a further fixed effect interaction term.

7 Safety Analyses

[Results will be presented by our partner ABBOTT.]

7.1 Extent of Exposure

The summary statistics will be produced in accordance with section 5.

7.2 Adverse Events and Serious Adverse Events

The safety data will be analyzed for the Safety Analysis Set.

Adverse events (AEs) will be coded with MedDRA, version 16.1. Only treatment emergent adverse events (TEAE) will be analyzed, i.e. AEs that started or worsened after start of study drug treatment. Treatment emergent AEs will be reported on a per-subject basis, i.e. counting subjects rather than events. This means that if a subject suffers the same AEs (i.e. assigned the same Preferred Term (PT)) repeatedly, the event will be counted only once. Repeated events per subject will be summarized according to the following rule: if a subject suffered the same AE more than once, the event will be assigned the worst severity, the closest relationship to the study drug and the earliest starting date. Both the TEAEs and the serious TEAEs will be summarized per primary System Organ Class (SOC), per Higher Level Term (HLT) by primary SOC and per PT by HLT and primary SOC. Severity and drug-event relationship of TEAEs are summarized separately. In the listings, however, all occurrences of an AE will be presented. Denominators will be based on the (size of the) Safety Set.

7.3 Clinical Laboratory Evaluations

The normal ranges differ between study centers.

Laboratory and vital signs values will be summarized by visit using the following summary statistics: the group mean and median value, the standard deviation, the range of the values, and the number of patients with a non-missing value. Both the absolute levels and the changes from baseline will be summarized, including changes from baseline to the last visit. Frequency tables will be presented for markedly abnormal values. Shift tables will be presented according to standard reference values, i.e., tables show the number of patients who are low, normal, or high at baseline and then at selected time intervals. Abnormal values will be identified in by-patient listings.

8 Summary of Changes to the Protocol

Changes to the statistical approach since the study was conceived:

- increase in sample size (resulting from a blinded sample size recalculation)
- higher dropout rate assumed (resulting in an increased number of patients to be allocated to the trial)
- change in the primary efficacy analysis: testing strategy was replaced by a modelling approach

8.1 Blinded sample size recalculation

For details see section 3.

8.2 Change in primary efficacy analysis

The primary efficacy analysis described in the protocol was changed. Instead of a non-parametric test (Kruskal-Wallis test followed by pairwise Wilcoxon Mann Whitney U-tests using a closed testing procedure) a *model-based analysis* was applied. The model-based principal analysis specifies a target parameter of interest (i.e. the incidence rates), rather than being purely based on hypotheses testing.

One reason for changing the analysis strategy was a proportion of study dropouts being higher than expected during the planning stage. A modelling approach seems more suitable to deal with methodological challenges resulting from a high proportion of incomplete primary efficacy data (derived from patient diaries) and to deal with different individual observation times, in particular within the time period of primary interest (i.e. the last 3 months of the 9 months treatment period). A GLMM with an offset term properly accounts for a varying number of evaluable days in order to estimate the incidence rates for each treatment arm.

References

References for Section 0 (Introduction and medical background information)

1. Committee on Hearing and Equilibrium guidelines for the diagnosis and evaluation of therapy in Menière's disease. American Academy of Otolaryngology-Head and Neck Foundation, Inc. *Otolaryngol Head Neck Surg* 1995; 113(3):181-5.
2. Minor LB, Schessel DA, Carey JP. Meniere's disease. *Curr Opin Neurol* 2004; 17(1):9-16.
3. Strupp M, Glaser M, Karch C, Rettinger N, Dieterich M, Brandt T. [The most common form of dizziness in middle age: phobic postural vertigo]. *Nervenarzt* 2003; 74(10):911-4.
4. Peron DL, Kitamura K, Carniol PJ, Schuknecht HF. Clinical and experimental results with focused ultrasound. *Laryngoscope* 1983; 93(9):1217-21.
5. Paparella MM, Mancini F. Vestibular Meniere's disease. *Otolaryngol Head Neck Surg* 1985; 93(2):148-51.
6. Anderson JP, Harris JP. Impact of Meniere's disease on quality of life. *Otol Neurotol* 2001; 22(6):888-94.
7. Filipo R, Lazzari R, Barbara M, Franzese A, Petruzzellis MC. Psychologic evolution of patients with Meniere's disease in relation to therapy. *Am J Otol* 1988; 9(4):306-9.
8. Hallpike C, Cairns H. Observations on the pathology of Menière's syndrome. *J Laryngol Otol* 1938; 53:625-55.
9. Schuknecht HF. Meniere's disease: a correlation of symptomatology and pathology. *Laryngoscope* 1963; 73:651-65.
10. Schuknecht HF. Endolymphatic hydrops: can it be controlled? *Ann Otol Rhinol Laryngol* 1986; 95:36-9.
11. Anatoli-Candela F. The histopathology of Menière's disease. *Acta Otolaryngol Suppl* 1976; 340:5-42.
12. Thomsen J, Bretlau P. General conclusions. New York: Georg Thieme Verlag Stuttgart; 1986.
13. Valvassori GE, Dobben GD. Multidirectional and computerized tomography of the vestibular aqueduct in Meniere's disease. *Ann Otol Rhinol Laryngol* 1984; 93:547-50.
14. Albers FW, Van Weissenbruch R, Casselman JW. 3DFT-magnetic resonance imaging of the inner ear in Meniere's disease. *Acta Otolaryngol* 1994; 114(6):595-600.
15. Mark AS. Contrast-enhanced magnetic resonance imaging of the temporal bone. *Neuroimaging Clin N Am* 1994; 4(1):117-31.
16. Fitzgerald DC, Mark AS. Endolymphatic duct/sac enhancement on gadolinium magnetic resonance imaging of the inner ear: preliminary observations and case reports. *Am J Otol* 1996; 17(4):603-6.
17. Yoshino K, Ohashi T, Urushibata T, Kenmochi M, Akagi M. Antibodies of type II collagen and immune complexes in Meniere's disease. *Acta Otolaryngol* Suppl 1996; 522:79-85.
18. Rauch SD, San Martin JE, Moscicki RA, Bloch KJ. Serum antibodies against heat shock protein 70 in Meniere's disease. *Am J Otol* 1995; 16(5):648-52.
19. Schuknecht HF, Suzuka Y, Zimmermann C. Delayed endolymphatic hydrops and its relationship to Meniere's disease. *Ann Otol Rhinol Laryngol* 1990; 99(11):843-53.
20. Lee KS, Kimura RS. Ischemia of the endolymphatic sac. *Acta Otolaryngol* 1992; 112(4):658-66.
21. Jackson CG, Glasscock ME, 3rd, Davis WE, Hughes GB, Sismanis A. Medical management of Meniere's disease. *Ann Otol Rhinol Laryngol* 1981; 90(2 Pt 1):142-7.
22. Klockhoff I, Lindblom U. Meniere's disease and hydrochlorothiazide (Dichlotride)--a critical analysis of symptoms and therapeutic effects. *Acta Otolaryngol* 1967; 63(4):347-65.
23. van Deelen GW, Huizing EH. Use of a diuretic (Dyazide) in the treatment of Meniere's disease. A double-blind cross-over placebo-controlled study. *ORL J Otorhinolaryngol Relat* Spec 1986; 48(5):287-92.
24. Silverstein H, Isaacson JE, Olds MJ, Rowan PT, Rosenberg S. Dexamethasone inner ear perfusion for the treatment of Meniere's disease: a prospective, randomized, double-blind, crossover trial. *Am J Otol* 1998; 19(2):196-201.
25. Harner SG, Driscoll CL, Facer GW, Beatty CW, McDonald TJ. Long-term follow-up of transtympanic gentamicin for Meniere's syndrome. *Otol Neurotol* 2001; 22(2):210-4.
26. Glasscock ME, 3rd, Thedinger BA, Cueva RA, Jackson CG. An analysis of the retrolabyrinthine vs. the retrosigmoid vestibular nerve section. *Otolaryngol Head Neck Surg* 1991; 104(1):88-95.

References for Section 1 to 8

References concerning trial endpoints or questionnaires:

Strupp M, Hupert D, Frenzel C, Wagner J, Hahn A, Jahn K, Zingler VC, Mansmann U, Brandt T. Long-term prophylactic treatment of attacks of vertigo in Menière's disease − comparison of a high with a low dosage of betahistine in an open trial. *Acta Otolaryngol* 2008; 128(5):520-4.

Calvert M, Blazeby J, Altman DG, Revicki DA, Moher D, Brundage MD, CONSORT PRO Group. Reporting of patient-reported outcomes in randomized trials: The CONSORT PRO Extension. *JAMA* 2013; 309(8):814-822.

Cohen HS, Kimball KT. Development of the vestibular disorders activities of daily living scale. *Arch Otolaryngol Head Neck Surg* 2000; 126(7):881-7.

Goebel G, Hiller, W. Tinnitus-Fragebogen (TF): Standardinstrument zur Graduierung des Tinnitus-Schweregrades - Ergebnisse einer Multicenterstudie. *HNO* 1994; 42:166-172.

Hiller W, Goebe G. Rapid assessment of tinnitus-related psychological distress using the Mini-TQ; *Int J Audiol* 2004; 43(10):600-4.

Jacobson GP, Newman CW: The development of the Dizziness Handicap Inventory. *Arch Otolaryngol Head Neck Surg* 1990; 116(4):424-427.

Methodological and Statistical references:

Adrion C, Mansmann U. Bayesian model selection techniques as decision support for shaping a statistical analysis plan of a clinical trial: An example from a vertigo phase III study with longitudinal count data as primary endpoint. *BMC Medical Research Methodology* 2012; 12(1):137.
URL http://www.biomedcentral.com/1471-2288/12/137

Carpenter JR, Kenward MG. *Missing data in randomised controlled trials – a practical guide*. National Institute for Health Research, Birmingham, 2007. Publication RM03/JH17/MK.
Available at http://missingdata.lshtm.ac.uk/downloads/rm04_jh17_mk .

Carpenter JR, Roger JH, Kenward MG. Analysis of longitudinal trials with protocol deviation: a framework for relevant, accessible assumptions, and inference via multiple imputation. *J Biopharm Stat* 2013; 23(6):1352-71

Chen Q, Chen MH, Ohlssen D, Ibrahim JG. Bayesian modeling and inference for clinical trials with partial retrieved data following dropout. *Stat Med* 2013; 32(24): 4180–4195.

Craig H. Mallinckrodt, W. Scott Clark, Raymond J. Carroll, Geert Molenberghs. Assessing Response Profiles from incomplete longitudinal clinical trial data under regulatory considerations. *J Biopharm Stat* 2003; 13(2):179-90.

Keene ON. Intent-to-treat analysis in the presence of off-treatment or missing data. *Pharm Stat* 2011; 10(3):191-5.

Mallinckrodt C. *Preventing and Treating Missing Data in Longitudinal Clinical Trials*. Cambridge University Press 2013.

Mallinckrodt C, Roger J, Chuang-Stein C, Molenberghs G, *et al.* Recent Developments in the Prevention and Treatment of Missing Data. *Therapeutic Innovation & Regulatory Science* 2014; 48:68-80.

Mallinckrodt C, Lin Q, Lipkovich I, Molenberghs G. A structured approach to choosing estimands and estimators in longitudinal clinical trials. *Pharm Stat* 2012; 11(6):456-61.

National Research Council (NRC). *The Prevention and Treatment of Missing Data in Clinical Trials. Panel on Handling Missing Data in Clinical Trials*. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press 2010.

O'Kelly M, Ratitch B. *Clinical Trials with Missing Data: A Guide for Practitioners*. Wiley 2014.

Ratitch B, O'Kelly M, Tosiello R. Missing data in clinical trials: from clinical assumptions to statistical analysis using pattern mixture models. *Pharm Stat* 2013; 12(6):337-47.

R Core Team (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

StatXact for Windows. Statistical Software for Exact nonparametric inference. User Manual 1996.

White IR, Carpenter J, Horton NJ. Including all individuals is not enough: lessons for intention-to-treat analysis. *Clin Trials* 2012; 9(4):396-407.

APPENDIX I: SOP '*Diary Assessment*'

The SOP Diary Assessment (Version 1.2, May 2014; Authors: Fischer CS, Adrion C, Strupp M) is an official consensus document and part of this SAP.

APPENDIX II: Technical Details

The study database is stored in SAS (Unix Version 9.2, SAS Institute Inc., Cary, NC). All statistical analyses will be performed using the statistical software package **R** version 3.1.1 (www.R-project.org) or SAS.

Quality assurance measures:

A second review statistician will independently reproduce the primary efficacy analyses. The reviewing statistician will have an overview of the entire analyses and will explicitly check the code producing the treatment estimates, as well as any other pieces of code as desired.

9 Reporting Conventions

P-values ≥0.001 will be reported to 3 decimal places; *p*-values less than 0.001 will be reported as "<0.001".

The mean, standard deviation (SD), and any other statistics other than quantiles will be reported to one decimal place greater than the original data. Quantiles, such as median, or minimum and maximum will use the same number of decimal places as the original data. Estimated parameters, not on the same scale as raw observations (e.g. regression coefficients) will be reported to 3 significant figures.

10 Program code (R or SAS)

10.1 Trajectory plots

Individual trajectory plots for Menière attacks will be displayed, together with the conditional posterior means of the number of attacks depending upon fixed and random effects after fitting the NB GLMM in a Bayesian setting (primary analysis). The same color is used to indicate observations and model-based estimates of the same patient.

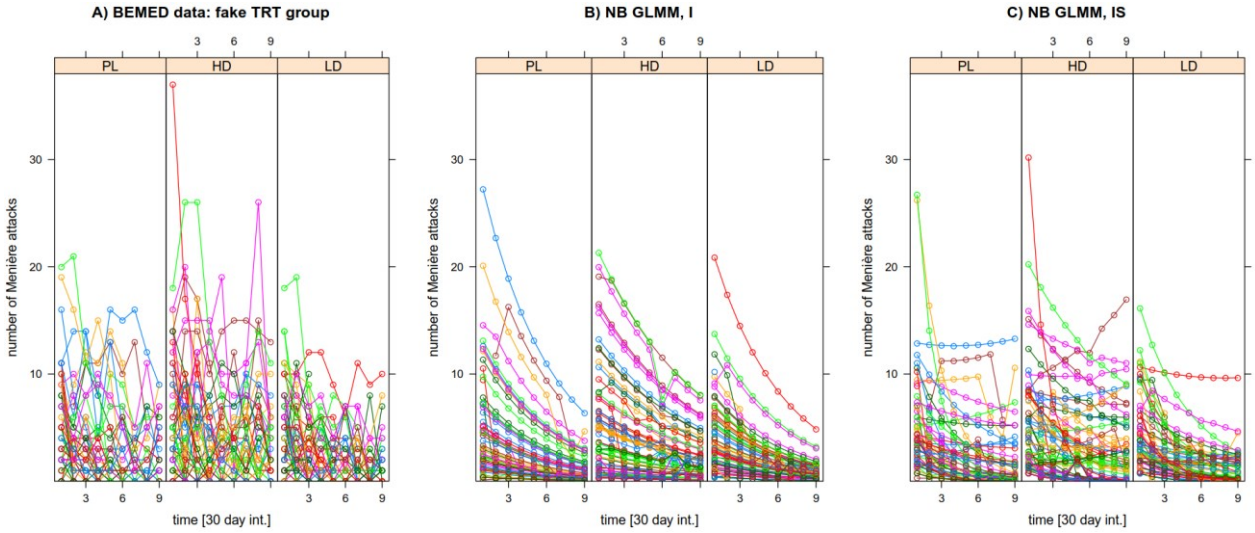

Figure 1 [Template] Trajectory plots for attack data. A) individual trajectories (without adjustment for the number of evaluated days). Figure B) and C) display the conditional posterior mean trajectories of the number of attacks depending upon fixed and random effects after fitting a NB GLMM with random intercepts (I), and with random intercepts and slopes (IS).

10.2 Data availability and completeness: Missingness Map

A summary of missing patterns will be obtained using PROC MI functionality in SAS:

SAS code fragment: Using PROC MI to examine patterns of missingness.

```
PROC MI data = attackdatawide NIMPUTE = 0; * dataset must be in wide format! ;
  VAR attacks; * variable 1 - variable9 ;
  ODS OUTPUT MISSPATTERN = mp_attacks;
RUN;
```

Chunck of R code to draw a missingness map:

```
require(Amelia)
missmap(attack.wide[, c(10:33)], rank.order = F, x.cex = 0.7)
```

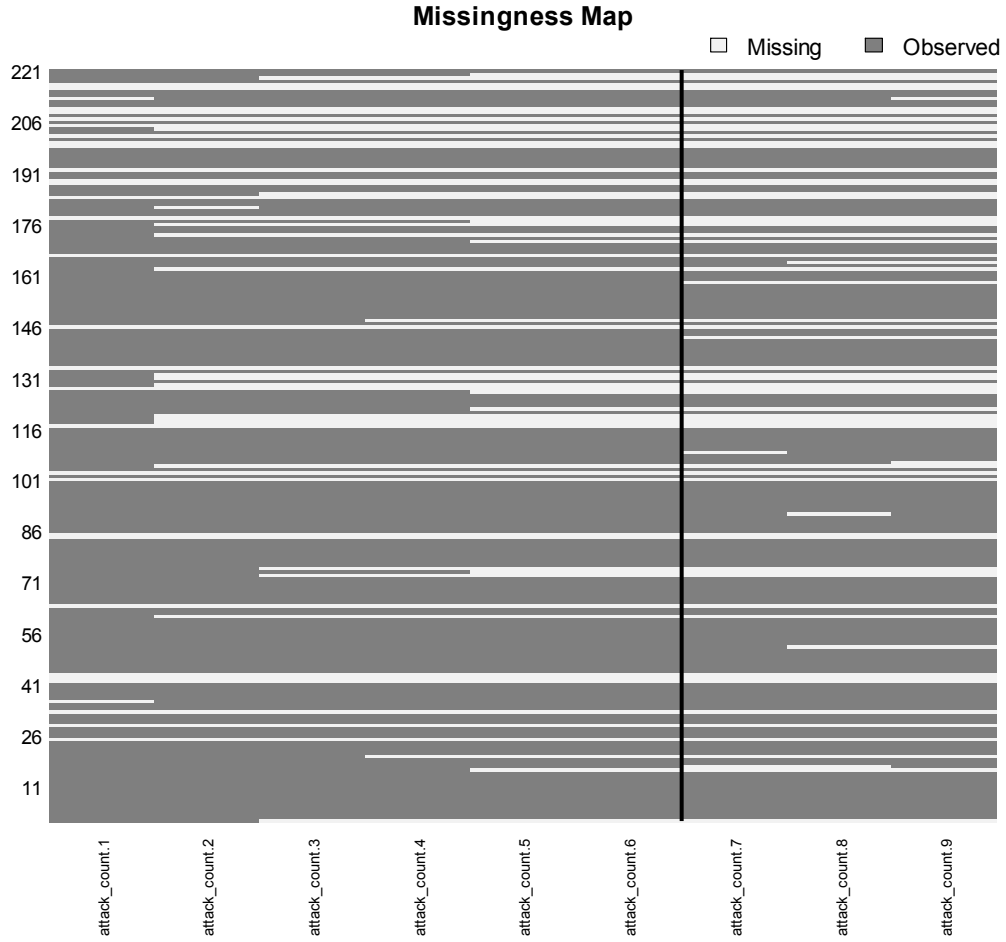

Figure 2 [Template] Heatmap showing where missingness occurs in the dataset within the primary time period of interest (interval 7, 8, 9). The figures displays monotone and intermittent missings of diary information. x-axis: number for each 30 day-intervals 1, 2, …, 9. y-axis: observation number representing PatIDs.

10.3 Negative Binomial GLMM

```
require(lme4)
library(lmerTest) # Tests for random and fixed effects for linear mixed effect models
```

The tests comprise type 3 and type 1 F tests for fixed effects, LRT tests for random effects, calculation of population means for fixed factors with confidence intervals and corresponding plots

```
# calculate least squares means for interaction TRT:timeint
# (timeint: time interval [unit: 30 days])
```

`lsmeans(.)`: Produces a Least Squares Means (population means) table with *p*-values and confidence intervals. The output resembles to what SAS software gives in a PROC MIXED statement. The approximation of degrees of freedom is Satterthwaite's.

```
mymodel = glmer.nb(...+ offset(log(eval.days)),
                   family = negative.binomial(link = "log"))
plot(lsmeans(mymodel))
```

```
lsmeans(mymodel, test.effs = "TRT:timeint")
          # test.effs: character vector specifying the names of terms to be tested.

plot(difflsmeans(m, test.effs = "TRT:timeint"))


library(glmmADMB)

### NB GLMM with Random slopes
model.1 <- glmmadmb(outcome ~ TRT*timeint + (timeint|id),
                                          data = daten,
                                          family = "nbinom",
                                          zeroInflation = FALSE,
                                          verbose = TRUE)

### Poisson lognormal GLMM with Random slopes
# all grouping variables must be factors:
ATTACKDATA$patid = factor(ATTACKDATA$patid)
ATTACKDATA$nobs = factor(ATTACKDATA$nobs)

fit.admb.1 = glmmadmb(attack_count ~  timeint + trt:timeint + offset(log(eval.days)) +
                       (timeint|patid) + (1|nobs),
                        data = ATTACKDATA[!is.na(ATTACKDATA$attack_count), ],
                        debug = TRUE,
                    mcmc = TRUE, # by default assuming flat, usually improper, priors...
                        # mcmc.opts = mcmcControl(mcmc = 50000))
                                                # (default chain of 1000 iterations)
                        zeroInflation = F, family = "poisson" )
                       )
```

10.4 Negative Binomial GLM

```
ATTACKDATA.789 = ATTACKDATA[ATTACKDATA$timeint %in% c(7,8,9), ]

# bei welchen Pt. gibt es Missings im Intervall {7,8,9} ?
require(nlme)

patid789 = gapply(ATTACKDATA.789, form = ~patid, which = "attack_count",
                  FUN = function(x) sum(!is.na(x)))
length(patid789)
table(patid789)


#--- aggregierten Datensatz generieren, der die Attackenfrequenz im Intervall {7,8,9}
#    aufsummiert => longitudinaler Aspekt entfernt:

ATTACKDATA.789.aggr = gsummary(ATTACKDATA.789[, -11], form = ~patid,
        FUN = function(x) sum(ifelse( sum(!is.na(x))> 0, sum(x, na.rm=T), NA)), inv=F )


glm.nb = glm.nb(attack_count ~  trt + offset(log(eval.days)),
                data = ATTACKDATA.789.aggr)

cbind(exp(coef(glm.nb)[2:3]), exp(confint(glm.nb)[c(2,3), ]))
```

10.5 WinBUGS and R-INLA code

10.5.1 WinBUGS code

Specification of the NB GLMM with bivariate random effects:

```
model
{
#   Longitudinal NB GLMM for BEMED using bivariate random effects
#
#
 for (i in 1:M)
      {
      y.obs[i] ~ dnegbin(p[i], r[i])
      logit(p[i]) <- (-1)*(a.0 + alpha[ID[i],1] +
                 (a.1 + alpha[ID[i],2] + a.2.2*treat.2[i] + a.2.3*treat.3[i]) * time[i])
       r[i] <- offset[i]
      }
#
      a.0 ~ dnorm(0, 0.001)
      a.1 ~ dnorm(0, 0.001)
      a.2.2 ~ dnorm(0, 0.001)
      a.2.3 ~ dnorm(0, 0.001)
#
      for (j in 1:N)
        {
        alpha[j, 1:2] ~ dmnorm(mu.0[], tau.mult[,])
        }
      tau.mult[1:2, 1:2] ~ dwish(R[,],2)
      sigma.mult[1:2, 1:2] <- inverse(tau.mult[,])
}
```

10.5.2 R-INLA code

For more details concerning the INLA approach see http://www.r-inla.org or Adrion & Mansmann (2012).

Chunk of R code to fit the NB GLMM using a Bayesian computational approach based on integrated nested Laplace approximations (INLAs):

```
# source("http://www.math.ntnu.no/inla/givemeINLA.R")
require(INLA)

#  NB model, Random Intercept + Random Slope
#-------------------------------------------

str(ATTACKDATA.bis9)
str(ATTACKDATA.bis9[!is.na(ATTACKDATA.bis9$attack_count), ])

length(unique(ATTACKDATA.bis9$patid))     # 221

n.block = max(ATTACKDATA.bis9$patid)

ATTACKDATA.bis9$i.intercept = ATTACKDATA.bis9$patid
ATTACKDATA.bis9$j.intercept = ATTACKDATA.bis9$patid + n.block  ## see doc for iid2d

formula.RIRS = attack_count ~ timeint + trt:timeint + offset(log(eval.days)) +
                      f(i.intercept,  model = "iid2d", n = 2*n.block) +
                      f(j.intercept, timeint, copy = "i.intercept")

# model specification (using an improved version of the grid integration...)

fit.inla.nb.IS <-
       inla(formula = formula.RIRS,
              data = ATTACKDATA.bis9[!is.na(ATTACKDATA.bis9$attack_count), ],
              family = "nbinomial",
              control.inla = list(strategy = "laplace", int.strategy = "grid",
                                  npoints = 21, diff.logdens = 4) ,
              control.compute = list(dic = TRUE, cpo = TRUE, mlik = T),
              control.predictor = list(compute = TRUE, cdf = c(.025, .975)),
                   # Prior setzen:
              control.fixed = list(cdf = c(0), prec.intercept = 0.001, prec = 0.001))
```

```
summary(fit.inla.nb.IS)
```

10.6 Definition of the *selected ear*

```
#  Inclusion-/Exclusion criteria:
#---------------------------------
eardat = read.table("d_inex.csv", header=T, sep =";",  dec = ".", na.strings = c("A","D","K", ""))

eardat = eardat[, c("patid", "incl2", "incl3", "incl4", "incl5", "incl6", "incl7")]
   # 0 = nein, 1 = ja

eardat$hoerverlust = ifelse( (eardat$incl3 !=1 | is.na(eardat$incl3)) & eardat$incl2 ==1 &
(eardat$incl4 !=1 | is.na(eardat$incl4)), "right",
                      ifelse( (eardat$incl2 !=1 | is.na(eardat$incl2)) & eardat$incl3 ==1 &
(eardat$incl4 !=1 | is.na(eardat$incl4)), "left", NA) )

eardat$hoerverlust.b = ifelse(eardat$incl4 == 1, "both", NA)

table(eardat$hoerverlust.b)  # 69 Pt. Hoerverlust=BOTH
table(eardat$hoerverlust)
# left right
#   71    80

eardat$hoerverlust = ifelse(!is.na(eardat$incl4) & eardat$incl4 == 1, "both", eardat$hoerverlust)
#

table(eardat$hoerverlust)
# both  left right
#   69    71    80     # Pt. 1046 hat keinen Hoerverlust


###############################################################################
#  Tinnitus ear
###############################################################################

eardat$tinnitusear = ifelse( (eardat$incl6 !=1 | is.na(eardat$incl6)) & eardat$incl5 ==1 &
(eardat$incl7 !=1 | is.na(eardat$incl7)), "right",
                      ifelse( (eardat$incl5 !=1 | is.na(eardat$incl5)) & eardat$incl6 ==1 &
(eardat$incl7 !=1 | is.na(eardat$incl7)), "left", NA) )

eardat$tinnitusear.b = ifelse(eardat$incl7 == 1, "both", NA)

table(eardat$tinnitusear.b, eardat$tinnitusear)
table(eardat$tinnitusear)

eardat$tinnitusear = ifelse(!is.na(eardat$incl7) & eardat$incl7 == 1, "both", eardat$tinnitusear)
#

addmargins(table(eardat$tinnitusear, eardat$hoerverlust))
            #hoerverlust
# tinn.| both left right Sum
#  both    30    7    11  48
#  left    26   63     1  90
#  right   13    0    68  81
#  Sum     69   70    80 219          # missing: Pt. 1046, 1039



###############################################################################
#    selected ear
###############################################################################


eardat$selectedear =
  ifelse(!is.na(eardat$hoerverlust) & eardat$hoerverlust %in% c("left", "right"),
       eardat$hoerverlust,
  ifelse(!is.na(eardat$hoerverlust) & eardat$hoerverlust == "both" & eardat$tinnitusear != "both",
  eardat$tinnitusear, NA) )

table(eardat$selectedear, useNA = "always")
addmargins(table(eardat$selectedear, eardat$hoerverlust, useNA = "always"))

eardat$hoerverlust = as.factor(eardat$hoerverlust)
```

```
eardat$tinnitusear = as.factor(eardat$tinnitusear)
eardat$selectedear = as.factor(eardat$selectedear)
###

set.seed(20140720)
random = rbinom(31, size=1, prob= 0.5)      # without Pt. 1046


# ears not classified yet:
id = which(is.na(eardat$selectedear))
length(id)     # 31

randomear = rep(NA, nrow(eardat))

randomear[id] = random
length(randomear)   # 221


table(randomear, useNA = "always")
#     0    1 <NA>
#    15   16  190

eardat$randomear = randomear
eardat$randomear = factor(eardat$randomear)
levels(eardat$randomear) = c("left", "right")       # defined codes: "0" = left, "1" = right


# finalize dataset, delete unnecessary auxiliary variables:
#-------------------------------------------------------------------------------

eardat$selectedear = ifelse(is.na(eardat$selectedear), eardat$randomear, eardat$selectedear)
eardat$selectedear = as.factor(eardat$selectedear)
levels(eardat$selectedear) = c("left", "right")

eardat = eardat[, c("patid", "hoerverlust", "tinnitusear", "selectedear")]
```

10.7 SAS Program to fill in missing dates for end of treatment

The following SAS program will be part of this SAP (author: Jos Nauta, ABBOTT Healthcare Products).

```
libname bemed "/home/clindata/lmu/bemed/data/database/derived";
options linesize=120 pagesize=35 pageno=1;


data date;
        set bemed.date;

        *** Define start date study drug medication;

        format start_dat DDMMYY10.;
        start_dat=einnahm1_dat;
        keep patid start_dat dov0 dov1 dov2 dov3 dov_lv th_von_dat th_end_dat;
run;



data einhalt;
        set bemed.einhalt (keep=patid visit therend_dat therend_d therend_m therend_y comm);
        length comment $60;
        comment=comm;
        drop comm;
        if therend_y ne .;
run;


data study_drug_dates;
        merge date einhalt;
        by patid;

        *** if visit=Ve and only stop day is missing then set day to 15;

        if visit='Ve' & therend_d in (.,.K) & therend_m >0 and therend_y > 0
        then therend_dat=mdy(therend_m,15,therend_y);
run;
```

```
data study_drug_dates;
        set study_drug_dates;

        *** Find latest stop date;

        by patid;
        if first.patid then nrows=0;
        nrows+1;
        if last.patid then lastrow=1;
run;


data study_drug_dates;
        set study_drug_dates;

        *** Assign study drug stop date;

        format stop_dat DDMMYY10.;

        if lastrow=1 & nrows=1 & visit='Ve' then stop_dat=therend_dat;
run;


*** CHECK;


proc print data=study_drug_dates;
        var patid dov0 dov1 dov2 dov3 dov_lv visit comment stop_dat;
        where stop_dat not in (.,.K) & comment ne ' ';
run;


proc print data=study_drug_dates;
        var patid dov0 dov1 dov2 dov3 dov_lv visit comment stop_dat;
        where stop_dat=.;
        by patid; pageby patid;
run;

/*
data study_drug_dates;
        set study_drug_dates;

        *** MANUAL IMPUTATIONS BASED ON INVESTIGATORS' COMMENTS;

        length stop_dat_comment $30;
   if patid=1042  then do; stop_dat=dov_lv;  stop_dat_comment='At last visit without study drug.';
                  stop_dat_imputed = 1; end;
        if patid=1051  then do; stop_dat=dov2;    stop_dat_comment='Stopped between V2 and V3.';
                       stop_dat_imputed = 1; end;
        if patid=1058  then do; stop_dat=dov3;    stop_dat_comment='Stopped after V3.';
                  stop_dat_imputed = 1; end;
        if patid=1067  then do; stop_dat=dov2;    stop_dat_comment='Study Drug Stopped between V2
                       and V3.';  stop_dat_imputed = 1; end;
        if patid=1073  then do; stop_dat=dov1;    stop_dat_comment='Study Drug Stopped between V1
                       and V2.';  stop_dat_imputed=1; end;
        if patid=3008  then do; stop_dat=dov0+1;  stop_dat_comment='Study Drug Stopped between V0
                       and T1.';  stop_dat_imputed=1; end;
        if patid=6003  then do; stop_dat=dov2;    stop_dat_comment='Study Drug Probably Stopped
                       after V2.';  stop_dat_imputed=1; end;
        if patid=14009 then do; stop_dat=dov2-1;  stop_dat_comment='Study Drug Stopped shortly
                       before V2.';  stop_dat_imputed=1; end;
        if patid=15002 then do; stop_dat=dov3+24; stop_dat_comment='Study Drug Stopped between V3
                       and V4.';  stop_dat_imputed=1; end;

        if lastrow;

        drop nrows lastrow visit comment;
run;
*/


proc print data=study_drug_dates;
        var patid dov0 start_dat dov1 dov2 dov3 dov_lv;
        where stop_dat in (.,.K);
run;
```

```
data study_drug_dates; set study_drug_dates;

        *** IMPUTATIONS;
        if stop_dat in (.,.K) then
        do;
                if start_dat not in (.,.A) then
                do;
                  if dov1=.  & dov2=.  & dov3=. & dov_lv=. then stop_dat= start_dat+1;
                  if dov1^=. & dov2=.  & dov3=. & dov_lv=. then stop_dat= int((start_dat+dov1)/2);
                  if dov1^=. & dov2^=. & dov3=. & dov_lv=. then stop_dat= int((dov1+dov2)/2);
                end;

                if      dov1^=. & dov2=.  & dov3^=. then stop_dat= int((dov2 + dov3)/2);

                if dov1^=. & dov2=.  & dov3=. & dov_lv^=.  then stop_dat= int((dov1+dov_lv)/2);
                if dov1^=. & dov2^=. & dov3=. & dov_lv^=.  then stop_dat= int((dov2+dov_lv)/2);
                if dov1^=. & dov2=.  & dov3^=.             then stop_dat= int((dov1+dov3)/2);
                if dov1^=. & dov2^=. & dov3^=. & dov_lv=.  then stop_dat= int((dov3+dov3)/2);
                if dov1^=. & dov2^=. & dov3^=. & dov_lv^=. then stop_dat= dov_lv;

                if stop_dat^=. then stop_dat_imputed=1;
        end;

        study_drug_duration = stop_dat - start_dat + 1;
run;



*** CHECK;


proc sort data=study_drug_dates;
  by study_drug_duration;
run;


proc print data=study_drug_dates;
        var patid dov0 start_dat dov1 dov2 dov3 dov_lv stop_dat study_drug_duration
stop_dat_imputed;
run;


proc print data=study_drug_dates;
        var patid dov0 dov1 dov2 dov3 dov_lv stop_dat therend_dat
            stop_dat_imputed stop_dat_comment;
        where therend_dat ^= stop_dat & therend_dat not in (.,.K);
run;
```

11 Date for Treatment end – patient-specific decisions

PATID	dov0	dov1	dov2	dov3	dov_lv	th_end_dat	start_dat	therend_y	therend_dat_IBE	stop_dat_Nauta	stop_dat_imputed	decision
3008	05.10.2010		03.03.2011	06.05.2011	11.08.2011	09.10.2010	05.10.2010	2010	**09.10.2010**	**06.10.2010**	**1**	IBE
1003	28.05.2008	27.06.2008				27.06.2008	29.05.2008	2008	**27.06.2008**	**12.06.2008**	**1**	IBE
1037	30.07.2010	13.09.2010				13.09.2010	14.08.2010	2010	**13.09.2010**	**29.08.2010**	**1**	IBE
8022	18.01.2012	02.03.2012				16.03.2012	01.02.2012	2012	**16.03.2012**	**16.02.2012**	**1**	IBE
16010	21.06.2011	22.07.2011				21.09.2011	21.06.2011	2011	**21.09.2011**	**06.07.2011**	**1**	IBE
11005	31.01.2011	09.03.2011				25.02.2011	01.02.2011	2011	**25.02.2011**	**19.02.2011**	**1**	IBE
1073	29.08.2011	28.09.2011	28.11.2011	27.02.2012	29.05.2012	26.10.2011	05.09.2011	2011	**26.10.2011**	**28.09.2011**	**1**	IBE
1045	03.11.2010	02.12.2010	28.01.2011			22.01.2011	04.11.2010	2011	**22.01.2011**	**30.12.2010**	**1**	IBE
4020	06.11.2012	04.12.2012	09.03.2013			08.04.2013	07.11.2012	2013	**08.04.2013**	**20.01.2013**	**1**	IBE
1064	18.04.2011	18.05.2011	22.08.2011			14.11.2011	19.04.2011	2011	**14.11.2011**	**05.07.2011**	**1**	IBE
1002	07.05.2008	05.06.2008			21.10.2008	01.08.2008	08.05.2008	2008	**01.08.2008**	**13.08.2008**	**1**	IBE
1051	19.11.2010	15.12.2010	09.03.2011	18.05.2011		31.01.2011	20.11.2010	2011	**31.01.2011**	**09.03.2011**	**1**	IBE
1067	01.06.2011	01.07.2011	07.10.2011	29.11.2011	05.03.2012	07.10.2011	02.06.2011	2011	**07.10.2011**	**07.10.2011**	1	
6003	27.06.2008	31.07.2008	05.11.2008	22.12.2008	01.04.2009	01.04.2009	28.06.2008	2009	**01.04.2009**	**05.11.2008**	**1**	IBE
14009	20.05.2011	20.06.2011	29.09.2011	25.11.2011	27.02.2012	28.09.2011	20.05.2011	2011	**28.09.2011**	**28.09.2011**	1	
12001	21.04.2011	23.05.2011	22.08.2011		24.10.2011	15.10.2011	22.04.2011	2011	**15.10.2011**	**22.09.2011**	**1**	IBE
1022	16.06.2009	24.07.2009	22.10.2009		21.12.2009	11.11.2009	17.06.2009	2009	**11.11.2009**	**21.11.2009**	**1**	IBE
1080	27.06.2012	26.07.2012	23.10.2012	07.12.2012		07.12.2012	28.06.2012	2012	**07.12.2012**	**07.12.2012**	1	
1058	01.03.2011	01.04.2011	29.06.2011	30.08.2011	25.11.2011	30.08.2011	01.03.2011	2011	**30.08.2011**	**30.08.2011**	1	
15002	01.04.2011	02.05.2011	09.08.2011	06.10.2011	04.01.2012	13.09.2011	02.04.2011	2011	**13.09.2011**	**30.10.2011**	**1**	IBE
14028	19.03.2012	18.04.2012	09.07.2012	10.09.2012	06.12.2012	10.05.2012	19.03.2012	2012	**10.05.2012**	**06.12.2012**	**1**	IBE
1026	17.11.2009	14.12.2009	17.03.2010	14.06.2010	11.08.2010	10.05.2010	18.11.2009	2010	**10.05.2010**	**11.08.2010**	**1**	IBE
1044	02.11.2010	02.12.2010	02.03.2011	02.05.2011	28.07.2011	28.07.2011	03.11.2010	2011	**28.07.2011**	**28.07.2011**	1	
4011	18.07.2011	22.08.2011	21.11.2011	17.01.2012	12.04.2012	28.12.2011	19.07.2011	2011	**28.12.2011**	**12.04.2012**	**1**	IBE
1042	27.10.2010	29.11.2010	24.02.2011	26.04.2011	25.07.2011	29.11.2010	28.10.2010	2010	**29.11.2010**	**25.07.2011**	**1**	IBE
1087	23.10.2012	23.11.2012	22.02.2013	22.04.2013	22.07.2013	22.07.2013	23.10.2012	2013	**22.07.2013**	**22.07.2013**	1	
6001	09.04.2008	07.05.2008	30.07.2008	15.10.2008	07.01.2009	07.01.2009	10.04.2008	2009	**07.01.2009**	**07.01.2009**	1	
6002	09.04.2008	07.05.2008	31.07.2008	06.10.2008	08.01.2009	13.12.2008	10.04.2008	2008	**13.12.2008**	**08.01.2009**	**1**	IBE
11001	06.12.2010	03.01.2011	20.04.2011	14.06.2011	06.09.2011	01.01.2011	07.12.2010	2011	**01.01.2011**	**06.09.2011**	**1**	IBE
6004	07.12.2010	07.01.2011	05.04.2011	09.06.2011	07.09.2011	07.09.2011	07.12.2010	2011	**07.09.2011**	**07.09.2011**	1	
6006	08.02.2012	08.03.2012	07.06.2012	08.08.2012	08.11.2012	21.03.2012	08.02.2012	2012	**21.03.2012**	**08.11.2012**	**1**	IBE
1085	05.10.2012	09.11.2012	01.02.2013	05.04.2013	08.07.2013	03.07.2013	05.10.2012	2013	**03.07.2013**	**08.07.2013**	**1**	IBE
1086	08.10.2012	08.11.2012	04.02.2013	04.04.2013	12.07.2013	12.07.2013	08.10.2012	2013	**12.07.2013**	**12.07.2013**	1	
8015	29.07.2011	26.08.2011	25.11.2011	20.01.2012	04.05.2012	28.04.2012	30.07.2011	2012	**28.04.2012**	**04.05.2012**	**1**	IBE
8008	09.09.2009	04.11.2009	04.02.2010	07.04.2010	14.07.2010	12.06.2010	07.10.2009	2010	**12.06.2010**	**14.07.2010**	**1**	
8009	04.01.2010	08.02.2010	14.05.2010	14.07.2010	21.10.2010	02.10.2010	05.01.2010	2010	**02.10.2010**	**21.10.2010**	**1**	IBE
8005	07.12.2009	12.01.2010	06.04.2010	22.07.2010	01.10.2010		09.12.2009	2010	**10.03.2010**	**01.10.2010**	**1**	

Figure 3 `therend_dat_IBE`: date for treatment end as documented in the original SAS database. `stop_dat_Nauta`: date for end of treatment resulting from programming ("imputation"), although a date is available in the database. Both dates were checked manually. In the case the imputed value was not valid in contrast to the date documented in the database, this decision is stated by `decision = IBE`. `stop_dat_imputed` means that a new date for end of treatment was generated (variable `stop_dat_Nauta`). `start_dat=einnahm1_dat`.

12 Full Analysis and Per Protocol Set: BDRM decisions

After the BDRM and prior to unblinding some decisions were made concerning the FAS and PP set on a per-subject basis (inclusion and exclusion criteria, evaluation of vertigo diary etc.).

The inclusion criteria "attack history" means that at least two attacks per months for at least three subsequent months before trial enrolment had to be documented.

patid	SAF	FAS.manuell	PP.manuell	Incl.crit./Excl.crit	*Ausschluss aus FAS?*	*Ausschluss aus PP ?*
1012	1	**1**	0		Tagebuch nicht bewertbar, aber relevante Information für sekundäre Endpunkte	
1017	1	**1**	0		Tagebuch nicht bewertbar, aber relevante Information für sekundäre Endpunkte	
1018	1	0	0		**Incl.crit. Attack history**	
1020	1	0	0		Incl.crit. Attack history	
1022	1	1	1			
1023	1	1	0			Medikation vertauscht
1046	1	**0**	0		keine M.Menière Diagnose	
1053	1	0	0		**Incl.crit. Attack history**	
1060	1	1	0			Kapseln geöffnet
3004	1	**1**	**1**	chronischer BPPV	*bleibt sowohl in FAS als auch PP (das Einschlusskrit. bleibt als verletzt)*	
3009	1	**0**	**0**		Incl.crit. Attack history	
4022	1	**1**	**0**			keinerlei Tagebuch vorhanden
10004	1	**1**	**1**		**Incl.crit. Attack history - fullfilled**	
11010	1	**0**	**0**		Incl.crit. Attack history	
11017	1	0	0		**Incl.crit. Attack history**	
14015	1	**1**	**1**		**Incl.crit. Attack history - fullfilled**	
14016	1	**1**	**1**		**Incl.crit. Attack history - fullfilled**	

BEMED Trial

SOP: Attackenbewertung der Tagebücher

Autoren der SOP	Fischer Carolin, Adrion Christine, Strupp Michael
TRIAL FULL TITLE (Acronym)	Medical treatment of Menière's disease with betahistine: a placebo-controlled, dose-finding study (*BEMED*)
EudraCT Nr.	2005-000752-32
ISRCTN Nr.	ISRCTN44359668
Prüfplan-Code	04T-617

Grundsätzliches

- Für jeden Patienten intraindividuelle Beurteilung jedes einzelnen Eintrags anhand der SOP.
- In seltenen, sehr komplexen Situationen (z.B. mehrere Tage mit Schwindel, Cluster mit Schwindel) kann die SOP nicht einwandfrei angewendet werden. In diesen Fällen erfolgt eine Besprechung des Falls im Team oder Blinded Data Review Meeting (BDRM). Attackenbewertungen, die nicht anhand der SOP durchgeführt werden konnten, werden gekennzeichnet (Notiz auf der bewerteten Tagebuch-Seite oder auf einem Analysebogen).

Preliminaries

1. ***Hierarchische Ordnung***: Für die dokumentierten Schwindelereignisse wird bei der Bewertung folgende hierarchische Ordnung zugrunde gelegt:

   Drehschwindel > Schwankschwindel > Gangunsicherheit > Benommenheit

2. ***Startuhrzeit der Attacken***

   Fehlen bei der Startuhrzeit die Minutenangaben oder sind diese unplausibel oder nicht lesbar, dann wird die Uhrzeit in der Datenbank auf die volle Stunde gesetzt (Uhrzeit:= Stunde:00)

3. ***Schwindelereignisse an unplausiblen Datumsangaben***

   Eintragungen werden als ungültig definiert, und Attacken somit nicht gewertet (z.B. 30.02.)

4. ***Definition "maximaler Bewertungszeitraum"***

   Der maximale Bewertungszeitraum der Tagebücher umfasst die individuelle Studiendauer. Gewertet werden Attacken ab dem 1.Tag *nach* Randomisierung bis einschließlich 1 Tag *vor* dem individuellen Studienende
   (Annahmen: Patient hat bei Einschlussvisite sowie Abschlussvisite keine Möglichkeit einer Attacken-Dokumentation. Baseline liegt in der Regel nach Random.dat.: Random.dat = min {Baseline, Random.dat., Therapiebeginn})

5. Tage mit Einträgen, die aufgrund der SOP nicht in einer *anrechenbaren* Schwindelattacke (welcher Qualität auch immer) münden, werden als *„attackenfreie Tage"* gerechnet; somit werden eingetragene (Dauer-)Symptome, wie z.B. „Tinnitus", „Druckgefühl im Ohr" und „Änderung des Hörvermögens", als nicht vorhanden betrachtet und für die weitere Bewertung ignoriert.

I. Schwindelart

Es sind verschiedene Schwindelarten möglich. Im Falle von Mehrfacheintragungen bei „Art“ zu einem identischen Zeitpunkt wird bei der Bewertung die Hierarchie berücksichtigt (D>S>G>B).

1. **Drehschwindel und Schwankschwindel** wird immer gerechnet, wenn

   a) eine klare und sinnvolle Startuhrzeit notiert ist

   b) die Dauer ≥ 2 ist

   weitere Symptome müssen nicht vermerkt sein.

2. **Gangunsicherheit** wird gerechnet, wenn

   a) eine klare und sinnvolle Startuhrzeit notiert ist

   b) die Dauer ≥ 2 ist

   c) am Tag danach keine anrechenbare Dreh- oder Schwankschwindelattacke erfolgte.

   d) Wenn es am Tag davor allerdings zu einer anrechenbaren Schwindelattacke (egal welcher Art) kam, dann muss mindestens eines der folgenden **Menière-typischen Begleitsymptome** an diesem Tag im Vergleich zum Vortag neu verzeichnet werden: ***Ohrdruck, Tinnitus, Hörveränderung,*** **Geräuschempfindlichkeit, wackelnde Bilder, Übelkeit, Erbrechen und Fallen.**

Angaben von Kopfschmerz, anderen Sehstörungen, Lichtempfindlichkeit, Lähmungen, Stand-/ Gangunsicherheit, Herzrasen und Atemnot, sind nicht als Menière-spezifisch zu werten.

3. **Benommenheitsschwindel** wird nur gerechnet, wenn

   a) eine klare und sinnvolle Startuhrzeit angegeben ist

   b) die Dauer ≥ 2 ist

   c) die Stärke ≥ 2 ist

   d) am Tag davor oder danach keine anrechenbare Dreh- oder Schwankschwindelattacke erfolgte

   e) mindestens eines der folgenden Menière-typische Begleitsymptome an diesem Tag neu verzeichnet werden: Ohrdruck, Tinnitus, Hörveränderung, Geräuschempfindlichkeit, wackelnde Bilder, Übelkeit, Erbrechen und Fallen.

Angaben von Kopfschmerz, anderen Sehstörungen, Lichtempfindlichkeit, Lähmungen, Stand-/ Gangunsicherheit, Herzrasen und Atemnot, sind nicht als Menière-spezifisch zu werten.

II. mehrere Schwindelattacken pro Tag werden gewertet; wenn

a) eine Startuhrzeit vorhanden ist

b) sie zeitlich plausibel erscheinen (Startzeit der unterschiedlichen Attacken, angegebene Dauer)

c) die jeweilige Dauer ≥ 2 ist

d) und es sich um Drehschwindel-Attacken handelt.

e) Sollten an einem Tag mehrere Schwindelarten „vermischt“ sein, wird nur die Drehschwindelattacke(n) gewertet, da oft ein (kurzes) zeitlich begrenztes Schwank/Benommenheitsgefühl oder auch Gangunsicherheit sowohl vor, als auch nach der eigentlichen Menière-Attacke auftreten kann.

f) Sollten an einem Tag mehrere Ereignisse mit Schwankschwindel oder auch Gangunsicherheit auftreten ohne Drehschwindel, wird die zeitlich als erste auftretende Attacke gewertet, unabhängig von der Dauer oder Stärke der an diesem Tag noch nachfolgenden Attacken gleicher Schwindelart.

g) Benommenheits-Attacken werden nur gerechnet, wenn

- eine klare und sinnvolle Startuhrzeit angegeben ist
- die Dauer ≥ 2 ist
- die Stärke ≥ 2 ist
- am Tag davor oder danach keine anrechenbare Dreh- oder Schwankschwindelattacke erfolgte
- mindestens eines der folgenden Menière-typischen Begleitsymptome an diesem Tag neu verzeichnet werden: Ohrdruck, Tinnitus, Hörveränderung, Geräuschempfindlichkeit, wackelnde Bilder, Übelkeit, Erbrechen und Fallen.

Es wird bei mehreren Benommenheits-Attacken pro Tag die zeitlich als erste auftretende Attacke gewertet, unabhängig von der Dauer oder Stärke der an diesem Tag noch nachfolgenden Attacken gleicher Schwindelart.

III. Fehlende Startuhrzeit

a) Schwank- und Drehschwindelattacken mit Dauer ≥ 2 ohne Menière-typische Beschwerden werden gewertet. Sollten > 2 Tage ohne Startuhrzeit (d.h. fehlende Stunde) aufeinander folgen mit der Dauer = 5, wird eine interne Begutachtung erfolgen.

b) Gangunsicherheit mit Dauer ≥ 2 und Angabe neu aufgetretener Menière-typischer Begleitsymptome (siehe I 2d) wird gerechnet. Bei Gangunsicherheit an > 2 aufeinanderfolgenden Tagen ohne Startuhrzeit, bei denen sich die Begleitsymptome nicht ändern, wird nur der erste Tag dieses Symptom-Clusters gerechnet. Eine Anrechnung erfolgt nicht, wenn am Tag nach der ersten Gangunsicherheit eine anrechenbare Dreh- oder Schwankschwindel-Attacke erfolgte.

c) Benommenheit ohne Startuhrzeit wird nicht als Attacke gerechnet.

IV. Schwindelattacken **ohne Art** werden gewertet, wenn

a) klare Startuhrzeit

b) die zeitliche Plausibilität stimmt

c) die Dauer ≥ 2 ist

d) die Stärke ≥ 2 ist

e) o.g. Menière-typische Begleitsymptome (I 2d) neu aufgetreten sind

f) wenn am Tag davor, am gleichen Tag oder am Tag danach keine anrechenbare Dreh- oder Schwankschwindelattacke auftritt.

Bei >2 aufeinanderfolgenden Tagen ohne Art sollte eine interne Begutachtung erfolgen.

V. Attacken mit **fehlender Stärke** werden gewertet, wenn

a) es sich um eine Dreh- oder Schwankschwindelattacke oder Gangunsicherheit handelt, eine klare Startuhrzeit und eine Dauer ≥ 2 vorhanden ist.

b) Benommenheitsschwindel ohne Angabe der Stärke (soll ≥ 2) wird nie gewertet

VI. Fehlende Dauer

Schwindelereignis, egal welcher Art, wird nicht gewertet

VII. Mindestens 2 aufeinanderfolgende Tage mit Schwindelereignissen und Dauer=5 (d.h. > 180 Minuten)

(sollten bei Dauer=5 verschiedene Schwindelqualitäten sich mischen oder parallel auftreten, greift auch Abschnitt II.)

a) Jede Dreh- oder Schwankschwindelattacke mit klarer Startuhrzeit (bei plausibler Startuhrzeit) wird gewertet, auch ohne Begleitsymptome. Somit werden aufeinanderfolgende D-/S-Tage nicht infrage gestellt.

b) Gangunsicherheit wird gewertet, wenn eine klare und plausible Startuhrzeit vorhanden ist, und mindestens eines der o.g. Menière-typischen Begleitbeschwerden an jedem Tag in dieser Phase im Vergleich zum Vortag neu hinzugekommen ist (I 2d).

c) Benommenheitsschwindel wird nur gerechnet, wenn eine klare und sinnvolle Startuhrzeit dokumentiert wurde, die Stärke ≥ 2 und mindestens eines der Menière-typischen Begleitsymptome im Vergleich zum Vortag neu hinzugekommen ist (siehe I 2d).

Gangunsicherheits- und Benommenheitsschwindel-Tage **nach oder vor einem Tag mit Dreh- oder Schwankschwindel** mit der Dauer=5 werden nicht gewertet, da es als „Nachhängen“ oder als "Ankündigung" zu werten ist.

Sollte ein Patient an > 2 aufeinanderfolgenden Tagen Gangunsicherheit oder Benommenheitsschwindel (Dauer=5) aufweisen, und/oder die Startuhrzeit immer identisch sein, so werden diese Tagebuchseiten im Team diskutiert.

VIII. Mindestens 2 aufeinanderfolgende Tage mit Gangunsicherheits-/ Benommenheits-Ereignissen und 2 ≤ Dauer < 5

Attacken werden gewertet, falls

- eine klare und sinnvolle Startuhrzeit vorhanden ist
- bei Benommenheitsschwindel die Stärke ≥ 2 ist
- bei Benommenheitsschwindel neben Dauer und Stärke auch mindestens ein Menière-typisches Symptom auftritt
- an jedem aufeinanderfolgenden Tag mit Benommenheitsschwindel oder Gangunsicherheit ein Menière-typisches Begleitsymptom neu hinzukommt.
- am Tag danach keine anrechenbare Schwank- oder Drehschwindelattacke kommt: in diesem Falle wird die Attacke an dem Tag vor der Schwank-/ Drehschwindelattacke nicht gerechnet.
- Bei Gangunsicherheit muss mindestens ein Menière-typisches Begleitsymptom neu hinzukommen, falls am Tag zuvor eine Schwank-/ Drehschwindelattacke gewertet wurde.
- Benommenheitsschwindel nach einem Tag mit Schwank-/ Drehschwindelattacke wird nicht gerechnet.

Werden Tage nicht gewertet (z.B. aufgrund fehlender neuer Menière-Symptome), dann wird dieser Tag als *„attackenfrei“* im Sinne der SOP interpretiert, d.h. evtl. dokumentierte Begleitsymptome werden nicht berücksichtigt. ➔ Der nachfolgende Tag wird somit als „Reset“-Zeitpunkt interpretiert (und Ereignisse, die mehr als 1 Tag in der Vergangenheit liegen, werden bei der Bewertung nicht berücksichtigt).

Ungenaue oder unklare Angaben

a) Sind Angaben bei Dauer oder Stärke ungenau, z.B. 1-2, so wird immer der größere Code verwendet (=> worst case-Prinzip).

b) Wurden für ein bestimmtes Schwindelereignis unterschiedliche Arten dokumentiert (z.B. D/B), wird für die Bewertung der Schwindel mit der höchsten Hierarchiestufe verwendet (**Drehschwindel > Schwankschwindel > Gangunsicherheit > Benommenheit**).
Es werden aber alle Schwindelarten in die Datenbank eingetragen.

c) Sollte die Vermutung entstehen, dass die Aufzeichnung eine Systematik beinhaltet

D B G	2 3 2

wird versucht, den Patienten zu kontaktieren und nachzufragen; wenn dieser klar das System erklären kann, wird eine Notiz auf dem Analysebogen angefertigt und dies gemäß der Angaben des Patienten konkretisiert. Ansonsten würden die Vorgehensweisen a) und b) greifen.

d) Falls für die Dauer die genaue Stunden- oder Minutenzahl dokumentiert wurde anstatt der vorgegebenen Kodierung, wird bei der Wertung einer Attacke entsprechend kodiert. Bei einer Angabe von "20 Minuten" wird mittels SOP entschieden, dass dies eine Einberechnung als einmaliges Ereignis mit Dauer="2" nach sich ziehen würde.

e) Notizen im Tagebuch durch Study Nurses oder Prüfärzte, die die vom Patienten dokumentierten Eintragungen als Attacken belegen oder widerlegen oder Erläuterungen liefern, sind zu berücksichtigen. Sollte es hierdurch aber zu starken Gegensätzen der normalerweise zu wertenden Attacken gemäß SOP kommen, würde der Fall im BDRM besprochen werden.

f) Falls Einträge in falsche Datumsspalten gemacht wurden und der Patient dies durch Korrektur der Datumsangabe (Überschreiben der Zahl) oder Pfeile in die entsprechende Spalte kenntlich gemacht hatte, wird dies berücksichtigt.

g) Sonstige nicht zulässige Kürzel, Zahlen, Zeichen oder handschriftliche Erklärungen werden bei der Bewertung ignoriert. Beispiele: x, >, <, →, „Hitzewallung"

h) Falls bei mehreren Einträgen bei einem Patienten die SOP nicht greifen kann, und somit die Gefahr besteht, dass mögliche Attacken nicht gewertet werden, erfolgt eine Besprechung des Falls im BDRM.

IX. "Missing pages", Hinzunahme externer Quellen

a) Falls in der Patientenakte Durchschläge von fehlenden Tagebuchseiten vorhanden sein sollten, dürfen diese kopiert und für die Auswertung verwendet werden.

b) Bei Blättern ohne Notiz „keine Schwindelattacke" sollte in der Patientenakte nachgelesen werden, ob Ärzte oder Study Nurses (z.B. bei Telefoninterviews) Notizen bezüglich der Attackenhäufigkeit gemacht hatten. Hierdurch sind Lücken in der Attacken-Dokumentation manchmal auffüllbar. Bei Hinzunahme externer Quellen erfolgt ein

schriftlicher Nachweis auf dem entsprechenden Auswertebogen (z.B. „fehlende Information anhand der Akte rekonstruiert“).

c) Unleserliche/ unklare Angaben oder fehlende Angaben:
Als letzter Schritt darf der Patient (zur Vermeidung unnötiger Datenverluste) kontaktiert werden. Dabei wird der Sachverhalt kurz geschildert (z.B. „nicht lesbar, ob D oder B“)
➔ Patient wird entsprechendes Tagebuchblatt zugeschickt, und dieser soll nach eigenem Ermessen korrigieren und die Korrektur unterschrieben zurückschicken. Wenn der Patient angibt, dass er ab Monat X keine Attacken mehr erlitten, daher auch keine Tagebuchseiten mehr ausgefüllt habe, so darf dies auf die Auswertebögen übertragen werden mit dem Hinweis „nach telefonischer Rücksprache mit dem Patienten“.

d) Wenn der Patient nicht erreichbar ist oder trotzdem keine Klarheit geschaffen werden kann, werden diese Tagebuchseiten intern besprochen oder – bei fehlenden Seiten – diese Monate als „missing“ gewertet.

X. Umgang mit „leeren“ Seiten, Definition attackenfreier Monate

Ein Tagebuchblatt, welches eindeutig einem Kalendermonat zugeordnet werden kann und entweder komplett leer, durchgestrichen oder gekennzeichnet ist mit eindeutigen Symbolen (z.B. Ø, 0) oder Kommentaren, wird als ***attackenfreier Monat*** interpretiert.

Bei „ND“ oder sonstigen Angaben, die nicht eindeutig als *attackenfrei* bewertet werden können, erfolgt eine Kontaktaufnahme mit dem zuständigen Prüfarzt des Zentrums zur weiteren Klärung.

Publikationsliste

Methodisch orientiert

Adrion C, Mansmann U. Bayesian model selection techniques as decision support for shaping a statistical analysis plan of a clinical trial: An example from a vertigo phase III study with longitudinal count data as primary endpoint. *BMC Medical Research Methodology* 2012; **12**(1):137.

Medizinische Journals

Feil K, Adrion C, Teufel J, Bösch S, Claassen C, Giordano I, Hengel H, Jacobi H, Klockgether T, Klopstock T, Nachbauer W, Schöls L, Stendel C, Uslar E, Warrenburg B, Berger I, Naumann I, Bayer O, Müller HH, Mansmann U, Strupp M. Effects of acetyl-DL-leucine on cerebellar ataxia (*ALCAT* trial): study protocol for a multicenter, multinational, randomized, double-blind, placebo-controlled, crossover phase III trial. *BMC Neurology* 2017; **17**(1):7.

Adrion C, Fischer CS, Wagner J, Gürkov R, Mansmann U, Strupp M; on behalf of the BEMED Investigators. Betahistine therapy in patients with Menière's disease: Primary results of a long-term, multicentre, double-blind, randomized, placebo-controlled, dose-defining trial of efficacy and safety (*BEMED* trial). *BMJ* 2016; **352**:h6816.

Bender A, Adrion C, Fischer L, Huber M, Jawny K, Straube A, Mansmann U. Long-term rehabilitation in patients with acquired brain injury: A randomized controlled trial of an intensive, participation-focused outpatient treatment program. [Langzeitrehabilitation von Patienten mit erworbenen Hirnschädigungen: Eine randomisierte kontrollierte Studie zu einem intensiven teilhabeorientierten ambulanten Therapieprogramm.] *Dtsch Arztebl Int* 2016; **113**(38):634–41.

Gross L, Theiss HD, Grabmaier U, Adrion C, Mansmann U, Sohn HY, Hoffmann E, Steinbeck G, Franz WM, Brenner C. Combined therapy with sitagliptin plus granulocyte-colony stimulating factor in patients with acute myocardial infarction – Long-term results of the *SITAGRAMI* trial. *International Journal of Cardiology* 2016; **215**:441–445.

Brenner C, Adrion C, Grabmaier U, Theisen D, von Ziegler F, Leber A, Becker A, Sohn HY, Hoffmann E, Mansmann U, Steinbeck G, Franz WM, Theiss HD. SITAgliptin plus GRanulocyte colony-stimulating factor in patients suffering from Acute Myocardial Infarction: A double-blind, randomized placebo-controlled trial of efficacy and safety (*SITAGRAMI* trial). *International Journal of Cardiology* 2016; **205**:23–30.

Hüfner K, Frenzel C, Kremmyda O, Adrion C, Bardins S, Glasauer S, Brandt T, Strupp M. Esophoria or esotropia in adulthood – a sign of cerebellar dysfunction? *Journal of Neurology* 2015; **262**(3):585–92.

Schwab F, Ingrisch M, Marcus R, Bamberg F, Hildebrand K, Adrion C, Gliemi C, Nikolaou K, Reiser M, Theisen D. Tracer kinetic modeling in myocardial perfusion quantification using magnetic resonance imaging. *Magnetic Resonance in Medicine* 2015; **73**(3):1206–15.

Neugebauer H, Adrion C, Glaser M, Strupp M. Long-term changes of central ocular motor signs in patients with vestibular migraine. *European Neurology* 2013; **69**(2):102–107.

Schniepp R, Wuehr M, Neuhaeusser M, Benecke A, Adrion C, Brandt T, Strupp M, Jahn K. 4-aminopyridine and cerebellar gait: a retrospective case series. *Journal of Neurology* 2012; **259**(11):2491–3.

Fritsche L, Fleckenstein M, Fiebig B, Schmitz-Valckenberg S, Bindewald-Wittich A, Keilhauer C, Renner A, Mackensen F, Mössner A, Pauleikhoff D, Adrion C, Mansmann U, Scholl H, Holz F, Weber B.
A subgroup of age-related macular degeneration is associated with mono-allelic sequence variants in the ABCA4 gene. *Investigative Ophthalmology & Visual Science* 2012; **53**(4):2112–2118.

Strupp M, Kalla R, Claassen J, Adrion C, Mansmann U, Klopstock T, Freilinger T, Neugebauer H, Spiegel R, Dichgans M, Lehmann-Horn F, Jurkat-Rott K, Brandt T, Jen J, Jahn K.
A randomized trial of 4-aminopyridine in EA2 and related familial episodic ataxias. *Neurology* 2011; **77**(3):269–275.

Lezius F, Adrion C, Mansmann U, Jahn K, Strupp M. High-dosage betahistine dihydrochloride between 288 and 480 mg/day in patients with severe Meniere's disease: a case series. *European Archives of Oto-Rhino-Laryngology* 2011; **268**(8):1237–1240.

Fleckenstein M, Schmitz-Valckenberg S, Adrion C, Visvalingam S, Göbel A, Mössner A, von Strachwitz C, Mackensen F, Pauleikhoff D, Wolf S, Mansmann U, Holz F. Progression of age-related geographic atrophy: role of the fellow eye. *Investigative Ophthalmology & Visual Science* 2011; **52**(9):6552–6557.

Fleckenstein M, Schmitz-Valckenberg S, Adrion C, Krämer I, Eter N, Helb H, Brinkmann C, Issa P, Mansmann U, Holz F. Tracking progression with spectral-domain optical coherence tomography in geographic atrophy caused by age-related macular degeneration. *Investigative Ophthalmology & Visual Science* 2010; **51**(8):3846–3852.

Fleckenstein M*, Adrion C*, Schmitz-Valckenberg S, Göbel A, Bindewald-Wittich A, Scholl H, Mansmann U, Holz F. Concordance of disease progression in bilateral geographic atrophy due to AMD. *Investigative Ophthalmology & Visual Science* 2010; **51**(2):637–642. *[geteilte Erstautorenschaft]

Brinkmann C, Adrion C, Mansmann U, Schmitz-Valckenberg S, Holz F. Klinische Merkmale, Progression und Risikofaktoren bei geographischer Atrophie. [Clinical characteristics, progression and risk factors of geographic atrophy]. *Der Ophthalmologe* 2010; **107**(11):999–1006.

Scholl H, Fleckenstein M, Fritsche L, Schmitz-Valckenberg S, Göbel A, Adrion C, Herold C, Keilhauer C, Mackensen F, Mössner A, Pauleikhoff D, Weinberger A, Mansmann U, Holz F, Becker T, Weber B. CFH, C3 and ARMS2 are significant risk loci for susceptibility but not for disease progression of geographic atrophy due to AMD. *PLoS One* 2009; **4**(10):e7418.

Hüfner K, Barresi D, Glaser M, Linn J, Adrion C, Mansmann U, Brandt T, Strupp M. Vestibular paroxysmia diagnostic features and medical treatment. *Neurology* 2008; **71**(13):1006–1014.

Mansmann U, Crispin A, Henschel V, Adrion C, Augustin V, Birkner B, Munte A. Epidemiology and quality control of 245 000 outpatient colonoscopies. *Dtsch Arztebl Int* 2008; **105**(24):434–40.

Letter to the Editor

Adrion C, Strupp M, Mansmann U. Lessons learned from a recent superiority trial on intratympanic injections in refractory unilateral Meniere's disease? Commentary on Patel et al. *BMJ* Rapid Response (electronic letter to the editor), 12 March 2017, `http://www.bmj.com/content/355/bmj.i6185/rr`

Conference Proceedings in Zusammenhang mit dieser Dissertation

Adrion C, Mansmann U. *Bayesian model selection using INLA with application to longitudinal count data.* LGM2012 – The Second Workshop on Bayesian Inference for Latent Gaussian Models with Applications. Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway. 30 May – 1 June 2012. Abstract und Poster.

Adrion C, Mansmann U. *Bayesian model evaluation for longitudinal count data in clinical trials: application to vertigo data.* GMDS & DGEpi Jahrestagung, Mainz, 26.–29.09.2011. Abstract und Poster.
German Medical Science GMS Publishing House, Düsseldorf 2011; DOI: 10.3205/11gmds098.
Available online:
`http://www.egms.de/static/en/meetings/gmds2011/11gmds098.shtml`

Adrion C, Müller HH, Mansmann U. *Statistical concepts for the primary efficacy analysis in vertigo trials.* Postersession im Rahmen des Meetings des IFB Scientific Advisory Board, 21.07.2011, Klinikum Großhadern, LMU München.

Adrion C, Mansmann U. *Decision support to predefine the analysis of a longitudinal count outcome in a RCT – Bayesian tools for model selection.* 30th Annual Conference of the International Society for Clinical Biostatistics (ISCB), 23–27 August 2009, Prague, Czech Republic. Abstract und Poster.

Adrion C. *Prädiktive Modellvalidierung mittels Proper Scoring Rules: Hintergrund und Anwendung.* Tagung der Arbeitsgruppe 'Bayes-Methodik' der Deutschen Region der Internationalen Biometrischen Gesellschaft. 05.12.2008, Mainz. Vortrag.

Adrion C, Rückinger S, Mansmann U. *Bayesian model diagnosis and model validation for longitudinal count data.* 29th Annual Conference of the International Society for Clinical Biostatistics (ISCB), 17–21 August 2008, Copenhagen, Denmark. Abstract und Poster.

Adrion C, Rückinger S, Mansmann U. *Generalized linear mixed models for counting processes.* 'LIFESTAT 2008 – Statistics and Life Sciences': 54. Biometrisches Kolloquium/ First Conference of the Central European Network, München, 10.–13. März 2008. Abstract und Poster. Abstract Volume ISBN 978-3-86541-266-9.

Eidesstattliche Versicherung

Adrion, Christine

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Thema

> *Moderne biostatistische Beiträge für Therapiestudien bei Schwindelsyndromen mit Tagebuch-basierten Attackendaten*

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

München, 15. März 2018

Christine Adrion