Universitätsklinikum Ulm
Klinik für Allgemein - und Viszeralchirurgie
Ärztliche Direktorin Prof. Dr. Doris Henne-Bruns

AG Versorgungsforschung
Leiter Prof. Dr. Franz Porzsolt

Progress in Evidence-Based Medicine by critical appraisal of Evidence-Based Medicine: The HOST Catalogue - Concordance of study Hypothesis, Objective, Statistics and Translation

Dissertation
zur Erlangung des Doktorgrades der Medizin
der medizinischen Fakultät der Universität Ulm

Meret Sophie Phlippen
Berlin
2020

Amtierender Dekan:

1. Berichterstatter:

2. Berichterstatter:

Tag der Promotion:

For my parents

Table of Contents

List of abbreviations

BMJ:	British Medical Journal
CONSORT:	Consolidated Standards of Reporting Trials
EBM:	Evidence Based Medicine
EG:	Exempli gratia; For example
ETC:	Et cetera
JAMA:	Journal of the American Medical Association
MJ:	Malaria Journal
NEJM:	New England Journal of Medicine
P:	Page
RCT:	Randomized Controlled Trial
VS:	Versus

1 Introduction

1.1 The history of the randomized controlled trial

The Randomized Controlled Trial (RCT) is the current gold standard of clinical trials for comparing a treatment to a control.
It was first performed for medical research in 1948 by Austin Bradford Hill testing the "Streptomycin treatment of pulmonary tuberculosis" [27, page 4582]. The original idea came from Sir Ronald Aylmer Fisher demonstrating it by the Lady Tasting Tea Experiment [12]. Since then randomization and the correct statistical analysis of study groups were recommended for the conduct of a clinical trial [4]. It is thought to equalize patient's characteristics as much as possible [4].
Over the time, critical voices have been raised [1; 23; 42]. It became clear that RCTs are not always of necessity the best study design [7; 34] and should be handled with care [22; 43]: Forms of bias were observed [22; 40] and a lack of attention for methodologic and statistical standards [42].
In the mid-1990s, the new idea of the Evidence Based Medicine (EBM) could no longer be ignored [10; 38]. Antiquated medical routines dominated the daily clinical practice due to a lack of applying discoveries in research. EBM's aim to ensure the best available treatment of a patient ought to be accomplished by utilizing the most recent, critically inspected, findings in research [36]. In addition, Consolidated Standards Of Reporting Trials (CONSORT) was launched to develop guidelines aimed for the best possible reporting in RCTs [41].
To this day, researchers and research groups are working on improvements of the best available study design. EBM still is a form of guideline for the best clinical practice and teaching. It has its benefits, but also negative consequences, such as a misuse by the industry or the limited applicability for multimorbid patients [15].
It was also shown that the influence of commercialization is one of the main reasons for a decrease in the quality of RCTs. The sponsoring of studies often has a negative effect on the outcome and increases bias [13; 14]. An example is the antidiabetic drug rosiglitazone. Its side effect of increased risk of myocardial infarction was swept under the carpet to pos-

sibly generate a better sales volume [46]. Another example is the controversy with calcium-channel antagonists supported by physicians and pharmaceutical industry and its doubtful safety [44].

One decade later almost identical critical comments were published in the literature [9, 16, 17].

The most newsworthy movement, at the moment, is coming from the Centre for evidence-based medicine (EBM) in Oxford. They detected inconsistencies in trials [24] and now aim to find a way to implement "the best available research evidence to clinical practice integrating the values of patients" [18].

A series by the Journal of Clinical Epidemiology highlights a different study design: the pragmatic trial. It can be seen as an addition to the RCT, without replacing it. It serves with extra information, should be considered as an equal partner to RCTs and completes the missing information. Pragmatic trials "evaluate relative effectiveness under conditions routinely encountered in clinical practice…" [29, p.13], whereas clinical trials "deliver data on efficacy and safety of treatments, yet often insufficiently inform physicians, policy makers, and other stakeholders how treatments will actually perform in real-world clinical practice." [29, p.13].

Recently it was stated in the British Journal of Cancer: "Many reports of health research omit important information needed to assess their methodological robustness and clinical relevance. Without clear and complete reporting, it is not possible to identify flaws or biases to reproduce successful interventions or use the findings in systematic reviews and meta-analysis." [26, p.619]

It is obvious that the conduct of an RCT needs clear rules: all of these above publications refer to a lack of quality of reporting. This lack may be related to methodologic standards, different forms of bias, such as commercialization or direct sponsoring. However, specific solutions to improve the quality were rarely recommended.

1.2 Aim of the project

The aim of the project was to identify indicators that detect quality differences in published Randomized Controlled Trials. We considered the quality of a publication high, when the study hypothesis, the objective of a study and the applied statistical tests were clearly stated, i.e. the study hypothesis was clearly stated as superiority or the hypothesis was not clearly stated and had to be derived out of the context. We regarded a published RCT of

high quality when the type-I-error and type-II-error was taken into account, the calculated sample size matched the recruited sample size and when it was clearly stated if the hypothesis could be confirmed or not. Second, we checked the intra-test concordance of the investigated variables of all clinical studies i.e. (1) the general study hypothesis, (2) the detailed objectives of the general hypothesis including the assessed end points , (3) the details of all applied statistical test related to the objectives of the general hypothesis, (4) the translation of all study results into plain language. In our opinion, a study of high quality would test i.e. a unidirectional study hypothesis by having a unidirectional study objective and applying a one-sided test, considering the type-I-error and type-II-error, recruiting the calculated sample size and clearly stating the results of the study.

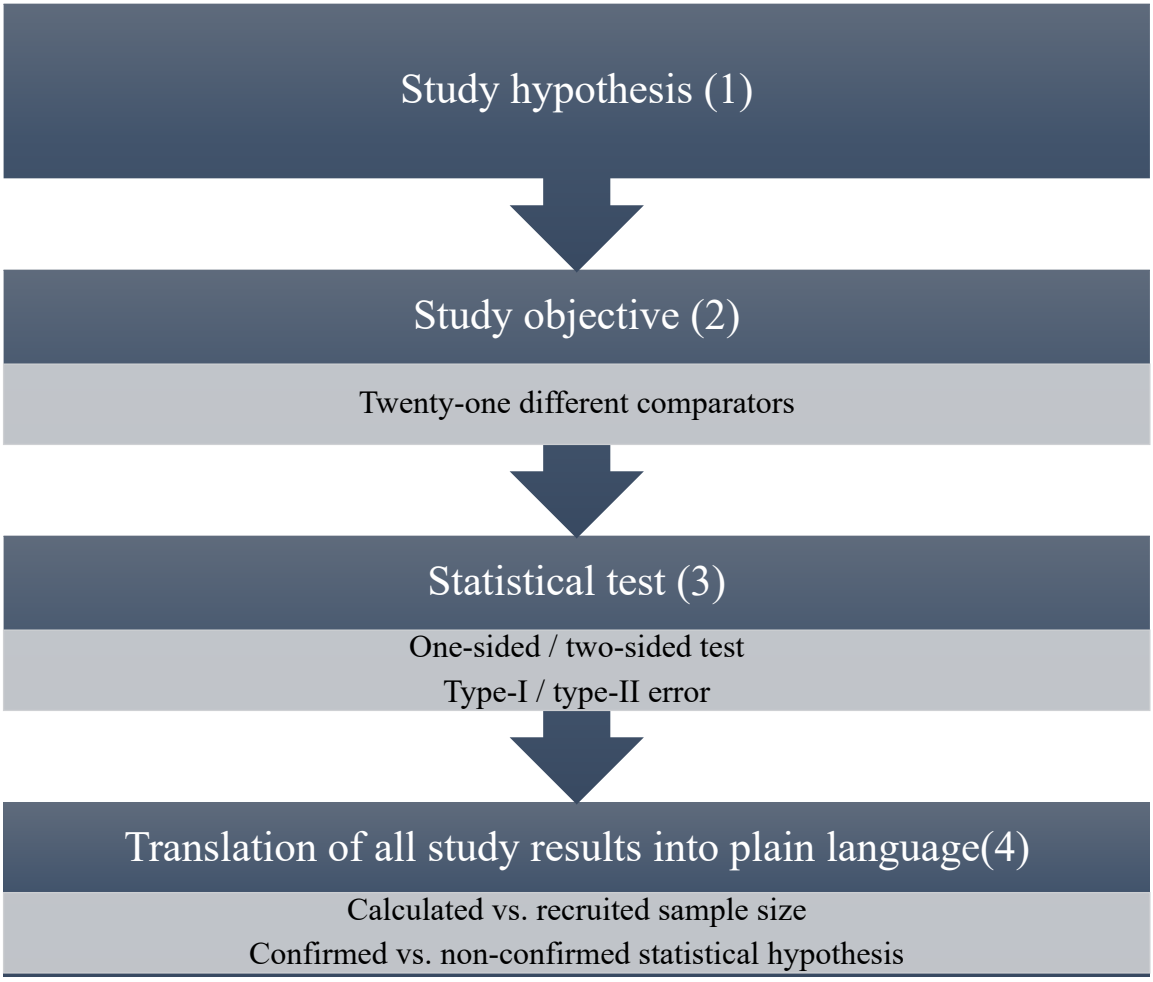

Figure 1.1: The six investigated variables and its concordance

Concordance of the six investigated variables: study hypothesis, its study objective with 21 different types of comparators, the applied statistical tests including one- or two-sided tests and the consideration of the type-I- and type-II-error, and the translation of all study results into plain language, which included the numbers for the calculated versus recruited sample size and the numbers for the studies with a confirmed or not confirmed hypotheses. It is the criteria we used to assess the RCTs.

The test for concordance is split into two parts. First, we checked for the concordance of a unidirectional or bidirectional hypothesis, unidirectional, bidirectional or multidirectional

objective, the application of one-sided and two-sided statistical tests. This check for concordance derives from one of the most important conditions for all trials. It is the precise definition of the study objective and the stated hypothesis, because all subsequent decisions such as the appropriate statistical test and translation of the study results depend on a precise study hypothesis.

Secondly the confirmation of the study hypothesis was correlated with a uni-,or bidirectional study hypothesis, with a uni-, bi-or multidirectional study objective, with one-sided or two-sided tests and with the concordance of type-I-or type-II-errors. These analyses are shown in table 1.1.

Table 1.1: Overview on the investigated variables and testing for concordance of some of these variables

Six variables were examined. Hypothesis: either clearly or not clearly stated and either uni- or bidirectional. Study objective: either uni-, bi- or multidirectional. Statistical testing: either one- or two-sided test and either the type-I- and type-II-error was considered or not. Study translation: comparison of the numbers for the calculated and the recruited sample size and if the hypothesis could be confirmed or not. All results for the six variables are shown separately in the results section and if necessary, its concordances were presented in detail in the stated chapters below.

	Hypothesis	**Study objective**	**Statistical test**		**Study translation of the results**	
	clearly, not clearly stated; uni- or bidirectional	uni-, bi- or multidirectional	one- or two-sided	type-I- or type-II-error or both	Sample size: Calculated vs. recruited	Confirmation or no confirmation of the study hypothesis
Hypothesis		See chapter 3.3.1	See chapter 3.3.1			See chapter 3.3.1
Study objective			See chapter 3.3.1			See chapter 3.3.1 and 3.3.2
One- or two-sided						See chapter 3.3.1
Type-I-/II-error						See chapter 3.3.3
Sample size						
Confirmation of study hypothesis						

1.2.1 Assumptions of the project

Our assumption is that the concordance of the study hypothesis with the study objective and the statistical tests, as mentioned in 1.2.1, is not as often in concordance as one would expect it to be. Therefore, we tested for concordance of these three variables in the second part of the results section (chapter 3.3).
In a well-performed study, a clearly stated unidirectional hypothesis (i.e. superiority) would match the unidirectional study objective (i.e. A vs. placebo), a one-sided test would be applied, the type-I- and type-II-error would be considered, the sample size would be achieved without a change in protocol and the hypothesis will be confirmed or not with statistically significant results.
Following this example, it can be concluded that the study objective and the statistical test have to fit the study hypothesis. Any inconsistency will jeopardize the study outcome.

In studies with inconsistencies we assume an influence on the study outcome. The study objective and the applied statistical test have to fit the study hypothesis. Absent or incorrect information on the interdependency of these variables will generate biased results. We cannot prove bias, but we can find indirect hints for bias in studies.
We did four analyses. Three were summarized in table 3.8, comparing the study hypothesis, the study objective, the statistical test and the rate of confirming the study hypothesis. The fourth screening was for studies considering the type-I-and type-II-error and their possibility of confirming their hypothesis. The second part of the results section states this inspection.

In particular we use the word ‘assumption’ for the description of our hypothesis to avoid confusion with the hypotheses of our investigated studies.

2 Materials and methods

The detailed information was derived from 120 published trials. They had to be randomized controlled trials published in medical journals.
20 selected papers each were included out of six different journals. The newest trials were selected chronologically until 20 trials were collected.
We worked in six teams for the selection and classification process.
Each of the co-authors was asked to select one journal she or he is reading consistently. The co-authors were Karthik Ghosh M.D., Ph.D. specialised in internal medicine and obstetrics and gynaecology, and Amit Ghosh M.D., Ph.D. from Rochester, specialised in internal medicine. They chose the New England Journal of Medicine (NEJM). The co-author G. Oscar Kamga Wambo M.D., Msc. from Berlin, specialised in internal medicine, public health, infectious disease and clinical economics, decided on analysing the Malaria Journal (MJ). Tania Gouvêa Thomaz M.D., specialised in physiology, pharmacology, biomedicine and Cristiane Moraes Ph.D., specialised in nephrology, cardiology, nutrition and dietetics, both from Niterói, selected Clinics as their journal. Valerio Balassone M.D. from Rome, a specialist in digestive surgery analysed the journal Annals of Surgery. Paola Rosati M.D. M.Sc., a paediatrician from Rome picked the journal Pediatrics. Franz Porzsolt M.D., Ph.D. and I, at this time a medical student from Ulm, elected the Journal of the American Medical Association (JAMA). Every team chose 20 RCTs chronologically beginning with the issue of November 2013 and then going backwards. We started the experiment in December 2013 and included the most current issues.

We included a variety of six different journals for a general overview of the current standard of randomized controlled trials. The difference between journals is measured by the journal`s impact factor, a calculated number of the influence of a journal. In our project the impact factor varies from 59.6 for NEJM [30], 37.7 for JAMA [25], 8.6 for Annals of Surgery [31], 5.5 for Pediatrics [48] and 3.1 for the Malaria Journal [28] to 1.2 for the journal Clinics [35]. These are the six journals we used for the project.

To detect potential weaknesses in the study hypothesis, objective, statistical test and translation of the study results of studies, we asked precise questions that had to be answered in our project. These questions were the following.

Each RCT was assessed by using six variables:

- Which type of the study hypothesis was applied? (superiority, non-inferiority, equivalence)
- Which study objective did exist? (describe exactly the compared study groups e.g. A vs. B, A vs. AB or specify if different)
- Which type of statistical test was used (e.g. one-sided, two-sided, both, none)
- Which type of the described error was considered? (e.g. type-I-error, type-II-error, both, none)
- Was there a difference in the calculated and the recruited sample size? (e.g. sample size reached as planned, sample size reached after change of protocol, sample size not reached due to… or specify if different)
- Was the study hypothesis confirmed? (e.g. superiority confirmed, superiority not confirmed, non-inferiority confirmed…)

Figure 2.1: Defined criteria used to assess the reporting of RCTs.
Detailed description of the six questions we used to assess the 119 studies. Each question was asked when reviewing the selected studies to sort the studies and find a statement on our assumption.

To evaluate the study hypothesis, we searched for the type of hypothesis of the trial.
The objectives of the study were reconstructed based on the selected experimental and control groups, the types of statistical tests (one- or- two-sided tests) and on the selected type of error (type I or type II error). We also recorded a potential difference in the calculated and recruited sample size and differences in the results of the statistical tests and the translation of the test results into plain language.

All results are documented in a table with a line for each paper and eight columns (see appendix).

The ethical commission decided on the 17th of May in 2017 that our project did not require any further consultation.

2.1 Trial-editing

The trials were downloaded as pdf-documents from the journal's website. The document had to be the full version of the trial, including the introduction, methods, results and discussion.

The selected articles were characterized by the journal's name, the year of the publication, the number of the issue and the first and last page of the published paper or the article number (see appendix).

2.2 Recording the six variables

2.2.1 Recording the type of hypothesis

Our first question identified the clear study hypothesis of the RCTs. The hypothesis can be unidirectional, i.e. testing for superiority or for non-inferiority. Alternatively, the hypothesis can be bidirectional, i.e. testing for equivalence.

The type of hypothesis of an RCT was recorded for each trial. It is reported in the introduction or methods section of the trial`s document. Hypotheses can also be described as the aim, expectation or the determination of a study.

There are different hypotheses to be found:

1. When a trial expects a treatment to perform better than its comparator the trial has a hypothesis of superiority. It is a unidirectional hypothesis.
2. If a trial wants to find a treatment being as good as or at least not worse than another treatment, the trial has a non-inferiority hypothesis. It is only testing in one direction and is therefore unidirectional.
3. Trials comparing two treatments and either hope to find one performing better or not worse than its comparator, conduct a trial with a hypothesis of superiority and non-inferiority. They are unidirectional hypotheses.
4. A trial evaluating the superiority for each treatment is conducting a trial with a hypothesis of superiority for each trial arm. It has a bidirectional hypothesis, considering each individual performed hypothesis and testing in both directions.
5. In a trial where no difference between two treatments is expected to be found, the hypothesis is testing for equivalence. The trial tests in two directions and aims to prove that neither a superiority nor an inferiority of study arms can be found. It tests a bidirectional hypothesis.

(see appendix (3rd column))

One example for a study stating their hypothesis of superiority is a paper out of the New England Journal of Medicine. It compared two study arms: one arm receiving an intervention (preventive percutaneous coronary intervention) and one arm receiving no intervention. The authors stated in the introduction that "the aim… was to determine whether pre-

ventive PCI… would reduce the combined incidence of death from cardiac causes, nonfatal myocardial infarction, or refractory angina." (see appendix (study 1, p. 1116)). A clear study hypothesis testing for superiority.
In addition, we examined if a non-inferiority testing was justified. An example for a trial testing an unjustified non-inferiority is from the malaria journal. It compares two malaria treatments. A justified testing for a non-inferiority requires an already documented sufficient effect of the medicament in comparison to placebo. The authors did not report on a previously completed placebo study and could consequently not comment on the justification of their study hypothesis (see appendix (study 36)).

In case of studies which do not explicitly report any study purposes in words, the hypothesis is to be derived out of the study context.
There are three different ways of concluding a hypothesis:

- It is either done by using the five different settings mentioned above and by searching for key words within the entire document. Key words are hypothesis, aim, expectation, determination, superior, inferior or equal and the corresponding verbs or subjects.
- A hypothesis can also be derived by looking at the study arms:
    - In case of placebo-controlled trials a superiority can directly be derived. The aim of placebo-controlled trials is to find the superiority of the compared treatment, not the non-inferiority or equivalence of it.
    - Another example are trials that compare treatment arms with an active treatment to no treatment. Here a hypothesis of superiority must be derived.
    - Trials comparing the same treatment, but with an additional treatment in one of the two study arms, want also to examine the superiority of the combined treatments.
    - If a trial compares two or more treatments and the superiority of one or the other is unknown, a trial is performing two or more tests for the superiority of each treatment and is therefore performing a bidirectional test.
- If the four options mentioned above do not apply to a trial, the hypothesis can be derived by reading the introduction. The current status of research and how the compared treatments performed in the past is described. Out of this description and the successful or less successful performances of the treatments, the expectation of

the outcome of a study can be derived. This can be a hypothesis of superiority, non-inferiority or equivalence.

(see appendix (3rd column in {brackets}))

An example for a derived hypothesis of superiority is a placebo-controlled trial about Dexmedetomidine being compared to the administration of saline (see appendix (study 58)). It was designed to evaluate their effect on shivering during spinal anaesthesia. Saline functioned as the control and was used as placebo. The authors did not explicitly state a hypothesis in words. A hypothesis for testing the superiority of Dexmedetomidine was therefore derived.

Another example, where a hypothesis for superiority was also derived, is a trial comparing study arms receiving treatment and no treatment (see appendix (study 50)). Patients either received 12 months L-thyroxine replacement or no treatment. There is no explicit hypothesis in words mentioned by the authors, but a hypothesis of superiority for the L-thyroxine replacement was derived from the study design.

A different study drew a comparison between usual care (= current standard care) and an intervention group (see appendix (study 87)). The usual care group received an information leaflet and the intervention group received five sessions of a family paediatrician-led motivational interview. It is obvious that the usual care group served as a control for a newly developed treatment protocol. The hypothesis must be derived as superiority for the intervention group.

An additional example for a derived hypothesis is study 61. The authors compared two operational techniques in bariatric surgery. One is considered the treatment of choice for the last 20 years, the other one is a new operative technique. The introduction states that “the evidence for the superiority of either surgical technique is still weak.” (see appendix (study 61, p. 690)). As no hypothesis was explicitly stated in words, we derived a testing of superiority for both treatment options. This study shows the difficulty of differentiating between a clearly stated hypothesis and a hypothesis not explicitly stated in words, but easy to comprehend and derive.

2.2.2 Recording the type of comparator

Our second aspect precisely describes the types of the twenty-one different experiments in the 119 RCTs. Examples for these experiments are A vs. no treatment or A vs. placebo or A vs. AB vs. AC vs. ABC. Each comparator of the experiment is described as a letter or symbol.

The type of comparator was recorded and used to classify the study objective. The study objective can be seen as the reason for the designing and completing of a study. At the beginning there is always a new idea of what should be investigated and why. Then it is reflected on how it can be investigated. The how is, which comparators are implemented and compared. We expect the comparators of a study to uncover the objective of a study.
A new drug is invented and should be implemented on the market. To test it for applicability, it is first compared to placebo. The comparators therefore are A vs. placebo. Its applicability is only affirmed when the new drug A is better than placebo. We can therefore state that the study objective is unidirectional, because the study does not want to test if placebo is better than the new drug A.
Studies with types of comparators such as A vs. placebo or A vs. no treatment were considered as unidirectional. A vs. B, otherwise, was regarded as a bidirectional study objective. Examples for multidirectional study objectives were A vs. B vs. C or AB vs. AC vs AD.

The type of comparator describes which comparison was performed, hence which treatments were compared.
The information was derived from the methods section, which gives detailed information about the treatments in the study arms.
To give a short oversight, treatments were coded in the form of the letter's "A", "B" or "C". Treatment arms receiving no intervention were indicated by the symbol "∅". "Placebo" stands for treatment arms receiving placebo.
It is additionally stated, if the treatment arms were new or old treatments. This information was collected by sampling the background in the introduction section. A treatment is new when only case studies exist, its data is scarce, or the impact of the new drug or therapy is yet unknown. An old treatment is an established standard treatment.
To understand how we summarized the comparators of a study, a study out of Clinics serves as an example (see appendix study 60)). The comparators are stated as A vs. AB and New vs. Old. The authors compared preoperative muscle training with usual care. The usual care group received no treatment preoperatively. All participants received the treatment A, open bariatric surgery. The effect of preoperative muscle training in obese patients undergoing open bariatric surgery has never been evaluated, it is a "new" option. Bariatric surgery is a standard procedure and an "old" intervention. Letter B stands for the intervention preoperative muscle training.

It was also distinguished if a treatment was compared to treatment or no treatment at all. No treatment could mean observation only (see appendix (study 5& 111)), watchful waiting (see appendix (study 8)) or participants in the treatment arm did not receive any feedback (see appendix (study 76)). In general, groups with no treatment or intervention served as a comparable control. All patients in study 15, for example, had a cardiac resynchronization therapy (CRT) implanted. However, only patients in the intervention group had their CRT turned on, but patients in the control group had their CRT turned off. According to the protocol of our project the experiment is A vs. ∅. Another example is study 92 (see appendix): Only one half of the infants received a foam plastic insert for their car seat. The other half of infants received no car seat insert and served as a control. The completed experiment is A vs. ∅.

2.2.3 Recording the type of statistical test

Our third point of interest was the selection and description of the applied statistical tests in the 119 RCTs. One variable was the exploration of the selection of the applied statistical test such as a one-sided test or two-sided test. A one-sided test covers only one of two possible options, for example A is better than no treatment. This one-sided hypothesis cannot confirm that treatment A is significantly worse than no treatment. The selection of a two-sided test is necessary when two possible options have to be analysed, for example A vs. B. A can be better than B, or A can be worse than B.

The type of statistical test was filtered out of the methods section and its subordinated section where the statistical analysis is described. In our study we focused on the use of one-sided or two-sided tests. In the literature -tailed has the same meaning as -sided.

A study in JAMA shows the difficulty of stating the type of statistical test. The study was designed to show the superiority of Otamixaban to heparin plus eptifibatide. In the methods section, however, both a one-sided with an 0.025 significance and a two-sided test with an 0.05 significance is stated. The one-sided test was used for the power calculation and the two-sided test for the comparison of the primary efficacy outcome (see appendix (study 115)).

2.2.4 Recording the type of described error

The other variable for the statistical tests of a study was the consideration of the risk of making a type-I-error and type-II-error. It was recorded for all of the selected 120 studies. This was done by reading through the methods section of every paper and especially scanning the description of the study analysis.

A type-I-error is the possibility of finding a difference between treatments, where no difference exists. It normally has a value of $\alpha = 0.05$ and can also be described as α-error, in the form of the hazard ratio and as the significance or confidence level in the literature. The type-II-error is calculated by $\beta = 1$-power and occurs when a really existing difference is not detected. Type-II-errors, or β-errors, are often described through the power of a study in the literature.

2.2.5 Recording the difference in the calculated and recruited sample size

Our fourth item was the assessment of the translation of the study results of each RCT respectively. We used two variables to screen the 119 RCTs. The difference in the calculated and recruited sample size and the type of statistical confirmation of a hypothesis.

We recorded the calculated number of patients needed and the number of patients being recruited in the study and checked, whether the calculated sample size of an RCT could be recruited at the end of study implementation.

If the calculated sample size was reached, we interpreted the power of a study as adequate. If it was not reached, we looked for reasons of unreached power in the paper. Indications could have been changes in the protocol, an early termination of a study or the study terminated without the completion of the calculated sample size.

Online databases such as *clinicaltrials.gov* or similar were additionally searched to find legitimate proof of a lack of a definition in protocol.

We valued interim analyses as a proof for a planned termination. Interim analyses define precise moments in the protocol were a committee examines the early results. They then can induce an early termination of the study, mostly due to safety reasons, and are stated in the paper or online.

If none of the above indications could be found for a reason of a non-achievement of the calculated sample size, we stated “not reached due to unknown reasons” (see appendix). When a calculated sample size is not stated, it is impossible to evaluate the power of a study.

An example for a non-achievement of the calculated sample size is a study in JAMA. Searching *clinicaltrails.gov* for study 105 (see appendix) showed that an early termination was not reported online and was not planned this way. Hence, it was stated that the calculated sample size could not be reached due to early termination.

2.2.6 Recording the type of statistical confirmation

The second variable for the assessment of the translation of the study results was, whether the study hypothesis of the RCT could be confirmed or not. For example, a study with a

hypothesis of superiority and statistically significant results, was a study with a confirmed superiority.
If a hypothesis could be confirmed or not can be seen in the results section of the trial`s document. In our project it is referred to the results of the primary study outcome. The results are evaluated by a significant or non-significant p-value.
If a study hypothesis could be confirmed, we state “confirmed”. If it was not confirmed, we state “not confirmed”.
Study 89, for example, compared neonates placed either in a plastic bag or not to test hypothermia one hour after birth. All neonates additionally received a standard thermoregulation protocol. Neonates in the plastic bag group had significantly lower rates of hypothermia compared to the other group without plastic bags. The p-value was 0.26. The neonates in the plastic bag group also had higher axillary temperature. The p-value was <0.001. The superiority of additionally placing neonates in plastic bags to prevent hypothermia is confirmed in this trial.

2.3 Recording the concordance of variables

In clinical trials some of these variables have to be concordant, e.g. a unidirectional hypothesis should be combined with a one-sided test. A two-sided test is not suitable for a unidirectional hypothesis. However, a bidirectional hypothesis definitely requires a two-sided test. The concordance of the type of hypothesis and the type of the statistical test (one- or two-sided) was therefore reviewed in the studies (see chapter 3.3.1).
Another example is the concordance of the study hypothesis and the study objective. A unidirectional hypothesis (A is superior to placebo, but not vice versa) should maintain a unidirectional study objective (A vs. placebo or A vs. AB). A bidirectional hypothesis (A is equal to B) demands a bidirectional study objective (A vs. B or AB vs. AC) (see chapter 3.3.1).
The concordance of the study objective and the statistical test (one- or two-sided) was also recorded. As well as the study hypothesis and the statistical test, the study objective and the statistical test should be concordant. A unidirectional study objective (A vs. nothing or A vs. placebo) requests for a one-sided statistical test. A bi- or multidirectional study objective, such as AB vs. AC vs. AD, needs a two-sided test (see chapter 3.3.1).
All results are classified for each study and summarized in one table.

Apart from the three above stated concordances there are four other concordances tested. All four include the confirmation of the hypothesis as a variable. The possibility of confirming or not confirming the study hypothesis was recorded for the different types of the study hypothesis (uni- or bidirectional, see chapter 3.3.1) the different forms of the study objective (uni-, bi- or multidirectional, see chapter 3.3.1) and for the different types of statistical testing (one- or two-sided tests (chapter 3.3.1) and the type-I-and -II-error (chapter 3.3.3)).

3 Results

3.1 Preliminary note

The results section presents the summary of the records of six variables of the 119 RCTs and recaps the concordance of the study hypothesis, its objective, the applied statistical test and the translation of the study results.
The section is divided into two parts. First, the results for each of the six variables will be shown in chapter 3.2. In the second part 3.3, the concordance of the three variables (study hypothesis, objective, statistical test) are displayed and linked by given examples. The concordance of the confirmation of the hypothesis is also linked with three other variables such as the study objective, the statistical test and the consideration of the type-I- and type-II-error.

All detailed notations on every RCT were recorded in a table (see appendix).
I decided to put it in the appendix, as it is a table which is too complex for the results section.
During the work on the RCTs it was noticed that one of the 120 RCTs was not a randomized trial, but only a controlled trial. Hence the 49th trial needed to be excluded. 119 out of 120 RCTs remained for the analysis.

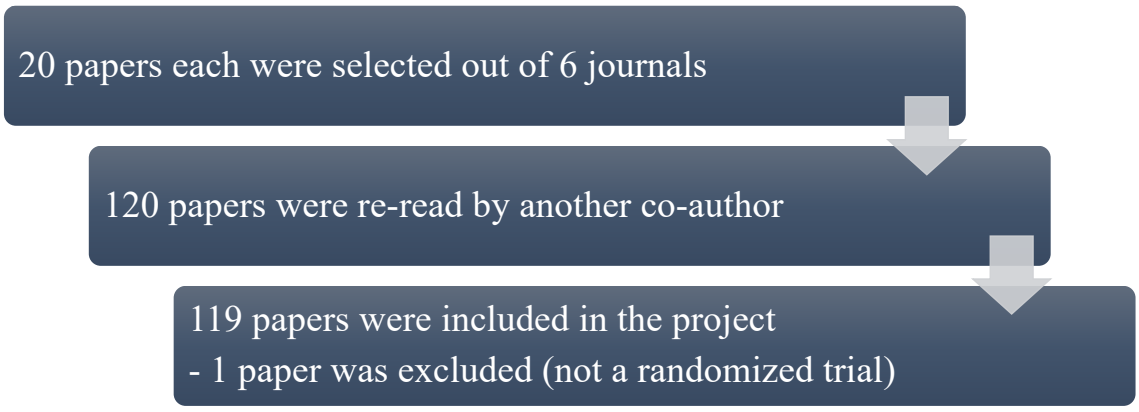

Figure 3.1: Identification of published papers of RCTs out of 6 different journals.
The boxes represent the three steps which were performed. The arrows show the chronological process.

3.2 The six variables

Here are the results concerning the six different variables we extracted from the 119 RCTs.

3.2.1 The type of hypothesis

In the following the occurrence of the hypotheses, either clearly stated or not clearly stated, are summarised. Results are shown for each journal and summarised for all 119 journals.

Table 3.1: Results in counting the appearance of described hypotheses in the 119 selected RCTs

Explanation:
In the column to the left, all types of hypotheses that were described in the RCTs are listed. Column 2 to 7 show the incidence for each journal. The last column sums up the numbers how often the hypothesis was described in all 119 RCTs. The order of options on the left is determined by the assumed significance from most to least in descending order. The last line shows the sum of each column.
NEJM: New England Journal of Medicine. Malaria: Malaria Journal. Ann Surg: Annals of Surgery. JAMA: Journal of the American Medical Association. Not stated: A hypothesis was not described in the RCT but could be derived. {}: Derived Hypothesis. Justified: The testing of a Non-Inferiority is justified due to prior testing or testing's. Not justified: The testing of a Non-Inferiority is not justified due to missing prior testing or testing's.

	NEJM	Malaria	Clinics	Ann Surg	Pediatrics	JAMA	Total
Superiority *	8	6	6	12	16	17	65
Superiority bidirectional *	1	0	0	0	0	0	1
Non-Inferiority (justified) *	3	7	1	0	0	0	11
Non-Inferiority (not justified) *	0	1	0	0	0	0	1
Superiority & Non-Inferiority (justified) *	3	0	0	0	0	0	3
Equivalence *	0	1	0	1	0	0	2
Not stated {Superiority} **	4	3	7	3	3	2	22
Not stated {Superiority bidirectional} **	1	1	4	3	0	1	10
Not stated {Equivalence} **	0	1	1	1	0	0	3
Not stated {Non-Inferiority, not justified} **	0	0	0	0	1	0	1
Hypothesis clearly stated *	**15**	**15**	**7**	**13**	**16**	**17**	**83**
Hypothesis not clearly stated **	**5**	**5**	**12**	**7**	**4**	**3**	**36**
Total	20	20	19	20	20	20	119

The most observed hypothesis was the testing of a superiority, including the not clearly stated hypotheses (73.1% (87 (65+22) out of 119 RCTs)).

The testing of a non-inferiority was observed either in combination with a superiority testing or alone. More RCTs tested a non-inferiority alone (10.1% (12 out of 119)) than in combination with a superiority testing (2.5% (3 out of 119 RCTs)). We also differentiated between a justified non-inferiority testing or a non-inferiority testing that was not justified. 87.5% of studies (14 out of 16 RCTs) with a non-inferiority testing were justified, whereas 12.5% (2 out of 16 RCTs) used a non-inferiority testing when not justified.

11 (9.24%) trials tested the superiority for more than one study arm, 10 out of 11 of the hypotheses had to be derived from the context. A minority of 4.2% (5 out of 119 RCTs) of the trials tested for equivalence between their study arms.
In 69.7% (83 out of 119 RCTs) a study hypothesis was clearly stated. In 30.3% (36 out of 119 RCTs) authors did not clearly state a study hypothesis and the hypothesis needed to be derived out of the context.

Most of the studies in JAMA and Pediatrics were superiority trials with a clearly stated hypothesis.
There is a difference in comparison to NEJM, the Malaria Journal and Clinics.
Three discriminations should be reported. More than half of the papers in Clinics did not state their hypothesis clearly in words. It is the paper with the highest number of papers with not clearly stated hypotheses.
The other observation is that one third of papers in the Malaria Journal tested a non-inferiority. No other paper did use the hypothesis as often as the Malaria Journal.
A study hypothesis of superiority and non-inferiority was only applied in NEJM.

3.2.2 The type of comparator

We included all possible study designs for RTCs. Our description of the study objective is tried to be as compact as possible. The study arms of each study are illustrated as single comparators and shown in table 3.2. The study objective is especially important in combination with the study hypothesis. We furthermore summarized the 21 different comparators into a unidirectional, bidirectional or multidirectional study objective. The results can be seen in the last lines of table 3.2. The table also shows, which study objectives were found, and which ones exactly are uni-, bi- or multidirectional.

Table 3.2: Results of summarizing the experiment completed of 119 RCTs

Explanation: The left column states the possible options describing the study experiment completed. Columns 2 to 7 show the incidence for each journal. On the right, the counted number of each option is given. The order of options on the left is determined by the assumed significance from most to least in descending order. The last line shows the sum of each column.
NEJM: New England Journal of Medicine. Malaria: Malaria Journal. Ann Surg: Annals of Surgery. JAMA: Journal of the American Medical Association. A, B, C, etc.: Treatment A, B, C, etc. New: New treatment. Old: Old or established treatment. ∅: No treatment. Other Forms of Experiments: All forms of experiments which appeared only once. *: Studies with a unidirectional study objective. **: Studies with a bidirectional study objective. ***: Studies with a multidirectional study objective.

	NEJM	Malaria	Clinics	Ann Surg	Pediatrics	JAMA	Total
A vs. B **	3	5	2	4	1	2	17
A vs. Placebo *	4	1	1	1	3	3	13
A vs. ∅ *	3	5	1	5	1	2	17
A vs. B (New vs. Old) *	4	5	7	8	7	5	36
A vs. AB (New vs. Old) *	3	1	1	1	5	3	14
A vs. B vs. C ***	0	1	1	0	0	0	2
AB vs. A + Placebo (New vs. Old) *	2	1	0	0	2	1	6
AB vs. AC (New vs. Old) *	1	0	0	0	0	0	1
A vs. AB *	0	1	0	0	0	0	1
A vs. B vs. C (New vs. Old) *	0	0	1	0	0	0	1
A vs. B vs. AB vs. AC vs. BC vs. ABC *	0	0	1	0	0	0	1
AB vs. AC vs. AD ***	0	0	1	0	0	0	1
A vs. B vs. C vs. AB vs. AC vs. BC vs. ABC vs. Placebo *	0	0	1	0	0	0	1
AA vs. AB vs. BC (New vs. Old) *	0	0	1	0	0	0	1
A vs. BB vs. BC (New vs. Old) *	0	0	1	0	0	0	1
A vs. B vs. AB *	0	0	0	1	0	0	1
A vs. B vs. Placebo (New vs. Old) *	0	0	0	0	1	0	1
A vs. B vs. AB vs. Placebo (New vs. Old) *	0	0	0	0	0	1	1
A vs. B vs. ∅ *	0	0	0	0	0	1	1
AB vs. Placebo, A vs. Placebo *	0	0	0	0	0	1	1
A vs. B vs. AB vs. ∅ (New vs. Old) *	0	0	0	0	0	1	1
Unidirectional study objective *	**17**	**14**	**15**	**16**	**19**	**18**	**99**
Bi-/multidirectional study objective ** / ***	**3**	**6**	**4**	**4**	**1**	**2**	**20**
Total	20	20	19	20	20	20	119

A majority of 44.5% of the studies (53 out of 119 RCTs) compared a treatment A with a treatment B. 67.9% of those (36 out of 53 RCTs) had a study objective, where one of the treatments is a new and the other one an older, respectively more established treatment.

The remaining 32.1% (17 out of 53 RCTs) compared two treatments being equal in terms of the establishment of one or the other trial.
To a lower extent, treatment A was compared with itself in a combination of an additional treatment AB in 12.6% (15 out of 119 RCTs) of the studies. 14 out of these 15 compared a new with an old treatment.
In general, more than half of the studies, 52.9% (63 out of 119 RCTs), tested a new treatment and compared it to an old one.
14.3% (17 out of 119 RCTs) of the studies compared a treatment A with no treatment.
Placebo-controlled trials were recorded in 16 % of the studies. Two thirds (13 out of 19 RCTs) of the placebo-controlled trials compared a treatment A with placebo. The other third (6 out of 19 RCTs) of study authors designed a study where the two groups received the same treatment, but one group received an additional treatment and the other group placebo as additional treatment.
For a better overview, experiments, which were recorded only once, can be seen in the last lines 9 to 22 in table 3.2. This case applied to 11.7% of trials (14 out of 119 RCTs).
An example is a study where the authors compared two different study groups (see appendix (study 112)). They gave one group either a combination of two medicaments or placebo and the other group either only one specific of the two medicaments or placebo. The experiment completed consequently is AB vs. placebo and A vs. placebo. None of the other 119 RCTs had the same experiment completed. Another study (A vs. B vs. C vs. AB vs. AC vs. BC vs. ABC vs. placebo) had eight different study arms (see appendix (study 43)). It is placebo-controlled and compares zinc, vitamin A and glutamine supplementation and combinations of it in study groups at the same time. In this case, the difficulty of selecting a one-sided or two-sided test can be shown.
Many of these studies, only once documented, (85.7% (12 out of 14 RCTs)) have three or more than three treatment arms and compare mostly a combination of treatments (85.7% (12 out of 14 RCTs)).
The difference between journals is not as apparent as it could be. NEJM, Pediatrics and JAMA published slightly more studies with placebo functioning as a control. It is, however, eye-catching that Clinics has the most studies with a multidirectional study objective.

To go further and to get an overview, we distributed the different types of comparators to two groups. One group consists of studies with comparators with a unidirectional study objective and the other with a bi-/ or multidirectional study objective. A unidirectional study

objective is interested in the result of a single direction. A bi-/ or multidirectional study objective looks at both directions of the results. The studies with a bi- and multidirectional study objective were summarized due to the fact that it does not make a difference in choosing the correct statistical test.

The distribution was done by classifying all studies with placebo as a comparator as unidirectional. All studies with ∅ (= no treatment) as one of the comparators was also categorized as unidirectional. Studies with a new treatment or a combined treatment, which were compared to an old or single treatment, were also grouped as unidirectional.

A bidirectional and multidirectional study objective can be seen in studies with balanced comparators, such as A vs. B or AB vs. AC vs. AD.

The quantity of studies with a unidirectional or bi-/multidirectional study objective can be found for each journal in the third and second last line of table 3.2.

82.2% (99 of 119) of the studies had a unidirectional study objective. In comparison 16.8% did form a bi- or multidirectional study objective.

3.2.3 Type of statistical test

The selection of the correct statistical test depends on the stated hypothesis, as mentioned before. It is therefore worth knowing which statistical test was used in the 119 RCTs and how often.

The chapter 3.3.1 of this results section will put the described statistical test in contrast to the described hypothesis and checks for concordance.

The table below shows the results for the numbers of one- or two-sided tests.

Table 3.3: Results in counting the use of statistical tests in terms of one-sided and two-sided tests in the 119 RCTs

Explanation:
The column to the left lists the options for the analysis of the RCTs. Columns 2 to 7 show the incidence for each journal. The last column sums up the counted times of the option. The order of options on the left is determined by the assumed significance from most to least in descending order. The last line shows the sum of each column.
NEJM: New England Journal of Medicine. Malaria: Malaria Journal. Ann Surg: Annals of Surgery. JAMA: Journal of the American Medical Association. Two-sided: A two-sided test was used. Not stated: Neither the use of a one-sided, nor a two-sided test was mentioned. One-sided: A one-sided test was used. One-sided & Two-sided: Both a one-sided and a two-sided test was used.

	NEJM	Malaria	Clinics	Ann Surg	Pediatrics	JAMA	Total
One-sided	1	4	1	1	0	1	8
Two-sided	12	8	6	10	13	19	68
One-& Two-sided (BOTH)	2	0	0	0	0	0	2
Not stated	5	8	12	9	7	0	41
Total	20	20	19	20	20	20	119

Table 3.3 shows that more than half of the RCTs (57.1% (68 out of 119 RCTs)) mentioned the use of a two-sided test. A minority of the RCTs (6.7% (8 out of 119 RCTs)) described a statistical testing in one direction. In one third of the RCTs (34.5% (41 out of 119 RCTs)) the use of a statistical test was not described. In 1.7% (2 out of 119) of the RCTs the use of both statistical tests was mentioned.

The analysis of differences between journals is noteworthy. Almost all the papers in JAMA stated the use of a two-sided testing, whereas more than half of the studies in Clinics and almost half of the studies in Annals of Surgery did not state any form of a statistical test. Over 50% of the papers in NEJM and Pediatrics did state the use of a two-sided test. Half of the one-sided tests were applied in the Malaria Journal.
All studies with a one- and two-sided testing scheme at the same time were published in NEJM.

3.2.4 Type of described error

It was interesting to find out how often a type-I-error and a type-II-error were considered for the statistical analysis. The synonyms for type-I-error, α-error, and type-II-error, β-error (see chapter 2.2.4), are not named explicitly and are summarized as type-I-error and type-II-error. Table 3.4 states the number of trials, which considered the type-I-error, type-II-error, both or none of the two errors.

Results for each journal are listed.

Table 3.4: Results in counting the tests of the type-I-error and the type-II-error in 119 selected RCTs

Explanation:
On the left possible options for the description of the type-I-error or type-II-error are listed. Columns 2 to 7 show the incidence for each journal. The last column sums up the counted RCTs with the corresponding mentions of type-I-error and type-II-error. The order of options on the left is determined by the assumed significance from most to least in descending order. The last line shows the sum of each column.
NEJM: New England Journal of Medicine. Malaria: Malaria Journal. Ann Surg: Annals of Surgery. JAMA: Journal of the American Medical Association. Both: both type-I-error and type-II-error were mentioned. Type-I-error: only the type-I-error was mentioned. Neither: neither the type-I-error, nor the type-II-error were mentioned. Type-II-error: only the type-II-error was mentioned.

	NEJM	Malaria	Clinics	Ann Surg	Pediatrics	JAMA	Total
Type-I-error	2	7	7	1	2	2	21
Type-II-error	1	0	0	0	1	0	2
Both (type-I- and type-II-error)	16	11	12	17	17	18	91
Neither	1	2	0	2	0	0	5
Total	20	20	19	20	20	20	119

More RCTs considered the type-I-error (94.1% (112 out of 119 RCTs)) than the type-II-error (78.2% (93 out of 119 RCTs)). A minority of trials did not comment on any of the errors (4.2% (5 out of 119 RCTs)).
76.5% (91 out of 119 RCTs) did both mention the type-I-error and type-II-error.
There was a difference between the Malaria Journal and Clinics and the four remaining journals. The two considered both of the errors not as often as the other journals. They therefore mentioned only the type-I-error and not the type-II-error. They are the two journals with the two lowest impact factors. JAMA has the highest number of journals regarding both errors.

3.2.5 The difference in the calculated and recruited sample size

For the examination of the power of a study the calculated sample size was chosen. It was examined in detail and a variety of six options were developed for the evaluation. Three out of six options describe the reasons for the non-achievement of the calculated sample size.

Results for the evaluation of the criteria can be found in table 3.5.

Table 3.5: Results in counting the options for defining the calculated sample size of the 119 RCTs

Explanation: The left column lists the eight different options for the evaluation of the calculated sample size. Columns 2 to 7 show the incidence for each journal. The right column gives numbers for the times an option was counted. The order of options on the left is determined by the assumed significance from most to least in descending order. The last line shows the sum of each column.
NEJM: New England Journal of Medicine. Malaria: Malaria Journal. Ann Surg: Annals of Surgery. JAMA: Journal of the American Medical Association. Reached as planned: Calculated sample size was reached as planned. Sample size not described: Calculated sample size was not described in the paper. Reached after change of protocol: Calculated sample size was reached after change of the study protocol. Not reached due to early termination: Calculated sample size could not be reached because the study was early terminated not accordingly to the study protocol. Early terminated according to protocol: Calculated sample size could not be reached, but study was terminated early according to the study protocol. Not reached due to unknown reasons: Calculated sample size could not be reached and for that no reasons could be found. Change of protocol& sample size not described: The study protocol was changed, and the sample size was not described in the paper. *: includes studies with the calculated sample size reached as planned and studies that terminated early according to protocol. **: includes studies that reached their calculated sample size after a change of protocol, which did not describe their sample size or which did not reach their sample size due to early termination or due to unknown reasons.

	NEJM	Malaria	Clinics	Ann Surg	Pediatrics	JAMA	Total
Reached as planned *	8	12	9	14	15	14	72
Reached after change of protocol **	5	2	0	2	0	1	10
Not reached due to early termination **	2	0	0	0	2	3	7
Not reached due to unknown reasons **	0	0	2	0	0	0	2
Sample size not described **	2	6	7	3	3	2	23
Early terminated according to protocol *	3	0	0	0	0	0	3
Change of protocol & Not reached due to unknown reasons **	0	0	0	1	0	0	1
Change of protocol & Sample size not described **	0	0	1	0	0	0	1
Number of recruited patients according to protocol *	11	12	9	14	15	14	75
Number of recruited patients not according to protocol **	9	8	10	6	5	6	44
Total	20	20	19	20	20	20	119

In 63% (75 out of 119 studies) the recruited sample size was according to protocol. 72 studies reached the sample size as planned, 3 studies terminated early according to the study protocol.

In 37% (44 out of 119 studies) the number of the recruited patients was not according to the study protocol. 10 studies reached their sample size after a change in the protocol. 7 studies did not reach their sample size due to an early termination which was not in line with the protocol found on *clinicaltrial.gov*. 2 studies stopped the study before reaching the calculated sample size without stating a reason. One study did both change its protocol and

did not reach its sample size without giving a reason and another study did change its protocol and did not state the calculated sample size.
In 20.2% (24 out of 119 RCTs) of the studies the calculated sample size was not stated at all.

The difference between journals can be seen by looking at studies with a reached sample size as planned. Around three quarters of the papers in Annals of Surgery, Pediatrics and JAMA could reach the calculated sample size as planned. 8 papers in the Malaria Journal, 9 papers in NEJM and 10 papers in Clinics represent 61.4% (27 of 44 RCTs) of the studies where the number of recruited patients does not accord to protocol. One quarter of the papers in NEJM had a change of protocol. About one third of the studies in MJ and Clinics did not state their calculated sample size at all.

3.2.6 Type of statistical confirmation of the hypothesis

If certain types of hypotheses are confirmed more or less often, will be seen in this chapter. It was recorded out of 119 RCTs, whether a trial could confirm its hypothesis or not. A statistically significant result of the primary outcome was our point of reference. Two different tables for the results were built. Table 3.6 lists the results for each paper and for every recorded type of hypothesis. Table 3.7 furthermore parts the confirmed or not confirmed hypotheses into correctly defined and not correctly defined hypotheses.

Table 3.6: Results in counting the 119 RCTs with significant or not significant results for the confirmation of their hypothesis

Explanation:
The left column lists the possible scenarios for confirming or not confirming the different hypotheses. Columns 2 to 7 show the incidence for each journal. The column to the right sums up the occurrence of the scenarios. The order of options on the left is determined by the assumed significance from most to least in descending order. The last line shows the sum of each column.
NEJM: New England Journal of Medicine. Malaria: Malaria Journal,.Ann Surg: Annals of Surgery. JAMA: Journal of the American Medical Association. Not stated: A hypothesis was not described in the RCT but could be derived. {}: Derived Hypothesis. Justified: The testing of a Non-Inferiority is justified due to prior testing or testings. Not justified: The testing of a Non-Inferiority is not justified due to missing prior testing or testings.

	NEJM	Malaria	Clinics	Ann Surg	Pediatrics	JAMA	Total
Superiority confirmed	4	2	1	7	6	4	24
Non-Inferiority (justified) confirmed	3	7	1	0	0	0	11
Non-Inferiority (not justified) confirmed	0	1	0	0	0	0	1
Non-Inferiority (justified) confirmed & Superiority not confirmed	3	0	0	0	0	0	3
Superiority not confirmed	5	4	5	5	10	13	42
Equivalence not confirmed	0	1	0	1	0	0	2
Not stated {Superiority} confirmed	5	2	3	1	1	1	13
Not stated {Equivalence} confirmed	0	1	1	1	0	0	3
Not stated {Superiority} not confirmed	0	2	8	5	2	2	19
Not stated {Non-Inferiority (not justified)} not confirmed	0	0	0	0	1	0	1
Total	20	20	19	20	20	20	119

About one third (36.4% (24 out of 66 RCTs)) of the superiority trials clearly stating their hypothesis could confirm their hypothesis. 40.6% (13 out of 32 RCTs) of the trials with a not clearly stated superiority hypothesis had a significant result concerning their study question.

Almost all of the trials with a non-inferiority hypothesis could confirm their assumption (12 out of 13). All of the trials with a justified non-inferiority hypothesis confirmed the non-inferiority. The non-inferiority trial which could not confirm its hypothesis is a trial out of the journal Pediatrics. Its hypothesis was not clearly stated, and it was derived that the study wanted to the test the non-inferiority of exothermic mattresses in preterm born infants to prevent hypothermia. Although many of the earlier conducted studies already showed results with many preterm born infants with temperatures out of the range, the

study was nevertheless conducted “to determine whether placing preterm newborns … results in more infants with rectal temperatures outside the range…” (see appendix study 99, p. 136).
None of the trials (2 out of 2) with a clearly stated equivalence-testing could confirm equivalent results between study arms. In contrast, all of the three studies with a derived equivalence-testing could confirm their hypothesis.
Trials with a superiority and non-inferiority testing could always confirm the non-inferiority of a treatment arm but never the superiority (3 out of 3 trials).

When differentiating between trials with a clearly stated and not clearly stated hypothesis there is a disproportion in the confirmed and not confirmed hypotheses. 47.0% (39 of 83) of the trials with a clearly stated hypothesis could confirm their hypothesis, whereas 44.4% (16 of 36) of the trials with a not clearly stated hypothesis could confirm their results.

Overall, 46.2% (55 out of 119 studies) of the studies could confirm their hypothesis and 53.8% (64 out of 119 studies) were not able to confirm their hypothesis.

A difference between journals can be demonstrated. When adding the different options for confirming a hypothesis, no matter which direction of hypothesis, there is a gap between NEJM and MJ and the four remaining journals. More than half of the papers in NEJM and MJ could confirm their hypothesis or one side of the hypothesis. It is the other way around with Clinics, Annals of Surgery, Pediatrics and JAMA. 13 (Clinics), 11 (Annals of Surgery), 13 (Pediatrics) and 15 (JAMA) of the papers could not confirm their hypothesis. 75% of the papers in JAMA could not confirm their hypothesis, whereas 75% of the NEJM papers were able to achieve significant results.

This analysis shows a difference in the ability of confirming certain hypotheses. It does not show the reasons for confirming or not confirming a hypothesis, it only shows the quantity of confirmed and not confirmed hypotheses.

Another analysis divides the types of hypotheses into two groups. The results of table 3.6 were used and the different types of hypotheses were divided into two groups of hypotheses. A ‘hypothesis defined correctly’ is understood as a clearly stated superiority hypothesis or a non-inferiority hypothesis that is justified. A non-inferiority hypothesis is justified,

when a superiority over placebo is already proven of both interventions. If the superiority over placebo has been proven of both interventions, a non-inferiority test is justified.
As 'hypothesis defined not correctly' we label studies in which an unjustified use of a non-inferiority hypothesis was detected and studies without a clearly stated hypothesis.
We wanted to see if there is a difference for studies with a correctly or not correctly defined hypothesis concerning their possibility of confirming their hypothesis or not.

Table 3.7: Results for studies with complete or incomplete documentation of the hypothesis and their rate of confirming results

Explanation:
The first and second column list the different options for the hypothesis of the studies and the first line lists the possible scenarios for confirming or not confirming the different hypotheses. The column to the right and the last line sum up the occurrence of the scenarios. The order of options on the left is determined by the assumed significance from most to least in descending order.
Justified: The testing of a Non-Inferiority is justified due to prior testing or testings. Not justified: The testing of a Non-Inferiority is not justified due to missing prior testing or testings. Not clearly stated: A hypothesis was not described in the RCT but could be derived. *: one study used an unjustified non-inferiority test and was therefore rated as 'hypothesis defined not correctly'.

	Hypothesis	**Confirmed**	**Not confirmed**	**Total**
Hypothesis defined correctly	Superiority clearly stated	24	42	66
	Non-Inferiority justified and clearly stated	11	0	11
	Non-Inferiority justified and Superiority clearly stated*	3	0	3
	Bidirectional hypothesis clearly stated	0	2	2
Hypothesis defined incorrectly	Not justified non-inferiority test	1	1	2
	Hypothesis not clearly stated	16	19	35
Total		55	64	119

46.3% (38 of 82) of the studies with a correctly defined hypothesis could confirm their hypothesis and 46% (17 of 37) of studies with a hypothesis defined incorrectly confirmed their hypothesis.

3.3 Necessary concordances of variables within studies

This chapter forms the second part of the results section. Now, the results for the combination of the criteria will be shown. Each chapter presents the concordance of two or more variables.

In RCTs are three variables that have to be concordant. The study hypothesis and the study objective have to be concordant. The study hypothesis and the statistical test need to be concordant and the study objective and the statistical test have to be concordant.

3.3.1 The type of the hypothesis, the study objective, the statistical test and the statistical confirmation of the hypothesis

The first test for concordance is a roundup of the three variables defining a study (study hypothesis, study objective and statistical test) and the variable of the study hypothesis being confirmed or not. The results can be seen below in table 3.8. The table gives an overview of the variety of studies and the different forms of conduct concerning the four variables.

The table represents an important aspect of the work. The incidence of the four variables were counted in the 119 studies and listed in 13 different compositions: Either a unidirectional hypothesis was stated with bidirectional study objective and a two-sided test and confirmed their study hypothesis. Or there was the composition in studies, where bidirectional study hypothesis was stated with a unidirectional study objective, without a statistical test stated and the hypothesis was not confirmed.

The table summarizes the different compositions of the four variables for all 119 studies. It was compiled to have a summary of the different forms of studies.

The study objective is linked to the hypothesis of a study. Either the study objective is clear in the first place and a hypothesis is evolved out of it, or the study hypothesis is already phrased, and the study objective becomes evident in the second place.

The statistical test is chosen according to the form of hypothesis. A one-sided test should be applied for a unidirectional hypothesis, e.g. superiority or non-inferiority. A two-sided test should be applied for a study with a bidirectional hypothesis, e.g. superiority & inferiority, superiority & non-inferiority or equivalence.

In order to the concordance of the statistical test and the hypothesis and the concordance of the hypothesis and the study objective, the statistical test and the study objective should also be concordant. A one-sided test should be used in studies with a unidirectional study objective. The other way around, a two-sided test should be used in studies with a bi- or multidirectional study objective.

Table 3.8: Concordance of three variables: the study hypothesis, its objective and the statistical test with the possibility of a confirmation of the hypothesis

Explanation:
The first line states the four variables, which are being compared. Beginning in line 2 and column 4 the results for the number of studies with the four variables are shown.
Unidirectional hypothesis: The author is interested in one side of the study outcome. Bidirectional hypothesis: The author is interested in both sides of the study outcome. Unidirectional study objective: A unidirectional study objective is interested in the result of a single direction. Bidirectional study objective: A bidirectional study objective looks at both directions of the results. Multidirectional study objective: A multidirectional study objective has multiple study arms, each having both directions as possible outcomes of a study. One-sided: A one-sided test was used. Two-sided: A two-sided test was used. One-sided & Two-sided: Both a one-sided and a two-sided test was used. Not stated: Neither the use of a one-sided, nor a two-sided test was mentioned. Not confirmed: Hypothesis could not be confirmed by statistically significant test. Confirmed: Hypothesis or one side of the hypothesis could be confirmed by a statistically significant test.
Grey background: Concordance of the study hypothesis, study objective and the statistical test.

Study hypothesis	**Study objective**	**Statistical test**	**Hypothesis**		**Total**
			Confirmed	Not confirmed	
Unidirectional	Unidirectional	One-sided	2	3	5
		Two-sided	27	29	56
		Not stated	15	18	33
		One-& two-sided	1	0	1
	Bi- or multidirectional	One-sided	2	0	2
		Two-sided	0	2	2
		Not stated	2	2	4
		One-& two-sided	1	0	1
Bidirectional	Unidirectional	Two-sided	1	2	3
		Not stated	0	1	1
	Bi- or multidirectional	One-sided	1	0	1
		Two-sided	2	5	7
		Not stated	0	3	3
Total			54	65	119

The study hypothesis was divided into two different forms of hypothesis, either they were unidirectional or bidirectional. Studies investigating superiority and non-inferiority were classified as unidirectional. Studies testing for the superiority of two options or testing for equivalence were categorized as bidirectional. In this chapter, it was not differentiated between clearly stated and not clearly stated hypotheses and justified or not justified non-inferiority tests.

The study objective was parted into uni- or bi- and multidirectional. The classification process is described in chapter 3.2.2.

Most of the studies tested a unidirectional hypothesis (104 of 119 studies, 87.4%) and the majority of these studies formed the matching unidirectional study objective (95 of 103 studies, 92.2%). Studies with a bidirectional hypothesis add up to 12.6% (15 of 119) of the studies. 26.7% (4 out of 15) of them formed a unidirectional study objective and 73.3% (11 out of 15) of them a matching bi-or multidirectional study objective. 12 of 119 studies (10.1%) did not have a concurring study hypothesis and study objective.
7 out of 67 studies (10.4%) with a unidirectional hypothesis and stated statistical test record to have been using a one-sided test. 58 out of 67 studies (86.6%) applied a two-sided test and tested a unidirectional hypothesis. We subtracted the studies without a stated statistical test in the calculation of the percentages, as we do not know which statistical test they used. 34.5% (41 out of 119 studies) studies did not state the use of a one-or two-sided test.
Studies with a bidirectional hypothesis used a two-sided test in 10 of 11 studies (90.9%) with a stated statistical test. One study (9.1% 1 out of 11 studies) applied a one-sided test for testing a bidirectional hypothesis.
5.1% (5 out of 99 studies) of the studies with a unidirectional study objective certainly used a one-sided test. Studies with a bi-or multidirectional study objective applied a two-sided test in 69.2% (9 out of 13 studies) of the studies.
Potentially 39 out of 99 studies with a unidirectional study objective (39.4%) could have been using a one-sided test, when we add the studies not stating a statistical test. Maybe 16 out of 20 studies (80%) with a bi-or multidirectional study objective did run the matching two-sided test, when we add the 7 studies not stating their statistical test.
Studies with a bidirectional study objective more often applied the appropriate statistical test.
Analysing the 119 studies, we can show, that in 41 studies the type of the used test was not stated at all. In 2 studies both tests, a one-sided and two-sided test were reported.
62 of 119 studies (52.1%) did not use the matching test to their study objective. This means 59 studies with a unidirectional objective used a two-sided testing scheme plus 3 studies with a bi-or multidirectional study objective had a one-sided testing scheme.

Details that are not explicitly stated in the work are that about one third of the studies have a superiority hypothesis and simultaneously applied a two-sided testing. Studies with a superiority hypothesis have a much lower incidence than non-inferiority studies for using a one-sided test: 5.6% vs. 36.4% (3 out of 54 studies vs. 4 out of 11 studies).

In terms of confirming the hypothesis there is an equal amount of studies with a unidirectional hypothesis confirming or not confirming their hypothesis: 50 studies vs. 54 studies. Whereas studies with a bidirectional hypothesis were less able to confirm their hypothesis: 4 studies vs. 11 studies.
Half of the studies with a unidirectional study hypothesis and study objective could confirm their hypothesis (47.4%, 45 of 95 studies). The studies with a unidirectional study objective and a bidirectional study hypothesis could only confirm the hypothesis in 1 out of 4 studies (25%).
Studies with a bi- or multidirectional study objective were able to confirm their hypothesis in about a third of the cases by having a bidirectional hypothesis (3 out of 11 (27.3%)) in comparison to having a unidirectional hypothesis (5 out of 9 (55.6%)).
Studies with a multidirectional study objective did confirm their hypothesis only in one study, having a unidirectional study hypothesis (see study 42, not shown in table 3.8).
More than half of the studies (4 out of 7 (57.1%)) with a unidirectional hypothesis and an applied one-sided test could confirm their hypothesis. Studies with a bidirectional hypothesis and an applied two-sided test could confirm their hypothesis in 42.9% (3 out of 7 studies) of the cases.
In studies with an unbalanced study conduct, which are studies with a unidirectional hypothesis and a two-sided test and studies with a bidirectional hypothesis and a one-sided test, 46.6% (27 out of 58 studies) vs. 100% (1 out of 1 study) could confirm their hypothesis.

The table above shows the results for all studies. But what happens, when we only include the studies with a stated statistical test and a clearly stated hypothesis?
83 studies (69.7%) clearly stated a study hypothesis. 78 studies (65.5%) stated a statistical test. 56 studies (47.5%) stated both their study hypothesis and their statistical test and 6 out of these 56 studies (10.7%) had a concordance of the study hypothesis, the study objective and the statistical test.
75 out of 83 studies (90.4%) had a concordant study hypothesis and study objective. 9 out of 56 studies (16.1%) showed a concordant study hypothesis and statistical test. 13 out of 78 studies (16.7%) had a study objective matching the statistical test.

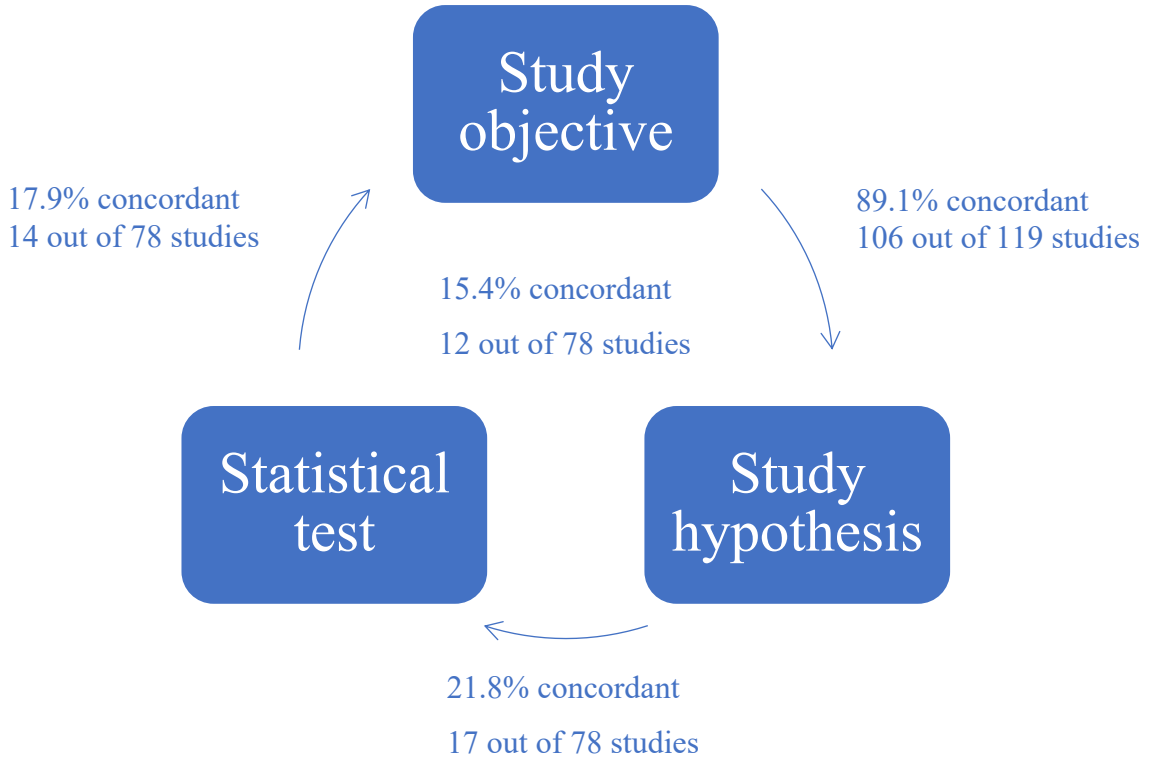

Figure 3.2: Rate of concordance of the study hypothesis, the study objective and the statistical test.
119 studies were analysed and included in the project. 78 of 119 studies stated a statistical test. Arrows show the concordances between two variables and the numbers below show the rate of concordance. The numbers in the middle of the figure represent the concordance of all three variables in the studies with a stated statistical test.

When we include all 78 studies with a stated statistical test and include all 119 studies, with the ones with a derived study hypothesis, for a demonstration of the study results we get the numbers shown in figure 3.2. The study objective and the study hypothesis were more or less concordant in the studies. A fifth of the studies showed a concordance of the study hypothesis and the statistical test. 17.9% of the studies had a concordant statistical test and study objective.

When going through the 78 studies with a stated statistical test a concordance of 15.4% of the studies can be counted. 12 studies had a concordant study objective, hypothesis and statistical testing.

3.3.2 The study with multiple types of comparators and the type of statistical confirmation of the hypothesis

The aim of this chapter is to see if the type of comparator has an influence on the ability to confirm a hypothesis or not. The type of comparator presents the conduct of a study, which

can be clear or rather chaotic. In this chapter it was given special attention to studies with multiple study arms.

A study with multiple comparators was considered to have a chaotic study objective.

The table below (3.9) shows the results for the possibility of confirming or not confirming their bidirectional hypotheses. It was interesting to see which hypotheses were applied in these complex study objectives and if they could be confirmed.

Table 3.9: Results for RCTs with ≥3 study arms concerning their hypothesis and its confirmation

Explanation:
The left column states the possible options for classifying the type of comparators. Columns 2 to 6 list the options for describing the confirmation of a certain hypothesis. Column 7 and Line 13 are summing up the total numbers.
Not confirmed: Hypothesis could not be confirmed by statistically significant test. Confirmed: Hypothesis could be confirmed by statistically significant test. Not stated: A hypothesis was not described in the RCT but could be derived. {}: Derived Hypothesis. A, B, C, etc.: Treatment A, B, C, etc. ∅: No treatment

	Superiority	**Superiority**	**Not stated {Superiority}**	**Not stated {Equivalence}**	**Not stated {Superiority}**	**Total**
	Superiority confirmed	Superiority not confirmed	Superiority confirmed	Equivalence confirmed	Superiority not confirmed	
A vs. B vs. C	0	1	0	0	2	3
AB vs. AC vs. AD	0	0	0	1	0	1
A vs. B vs. C vs. AB vs. AC vs. BC vs. ABC vs. Placebo	0	0	1	0	0	1
AA vs. AB vs. BC	0	1	0	0	0	1
A vs. BB vs. BC	0	0	0	0	1	1
A vs. B vs. AB vs. AC vs. BC vs. ABC	0	0	0	0	1	1
A vs. B vs. AB	0	1	0	0	0	1
A vs. B vs. Placebo	1	0	0	0	0	1
A vs. B vs. AB vs. Placebo	0	1	0	0	0	1
A vs. B vs. ∅	0	1	0	0	0	1
A vs. B vs. AB vs. ∅	0	1	0	0	0	1
Total	1	6	1	1	4	13

10 out of 13 studies (76.9%) could not confirm their hypothesis. Two studies could confirm the superiority of a treatment arm and one could confirm the equivalence of the study arms. Almost half of these studies (6 out of 13) did not clearly state a hypothesis and the hypothesis had to be derived from the context.

12 out of 13 studies tested for superiority. The remaining study tested for the equivalence of three study arms. The study tested three different combinations of topical glaucoma treatment. Every study arm tested timolol in combination with a prostaglandin analogue.

In comparison to studies with a unidirectional study objective, where 46.5% (46 out of 99 studies) could be confirmed or with a bidirectional study objective where 41.2% (7 out of 17 studies) could confirm their hypothesis, studies with a multidirectional study objective confirmed 14.3% (2 out of 14 studies) their hypothesis.

3.3.3 The type of described error and the type of the statistical confirmation of the hypothesis

Investigating those two variables, may show a concordance of a thoroughly detailed statement on the applied statistical methods in a study and the occurrence of confirming its hypothesis. The statement on the described error in the analyzed studies functions as a sort of quality control of the study outcome. We cannot test effects on the study outcome, but we can observe if the type-I-error was considered adequately in studies with a confirmed study hypothesis and if the type-II-error was considered enough in studies which could not confirm their hypothesis. Not reporting a potential error will increase the risk of not being aware of it.

Table 3.10 features the results for the selected RCTs.

Table 3.10: Results in sorting the 119 RCTs in terms of the confirmation of the hypothesis and considering the type-I-&-II-error

Explanation:
The left column states the two possible options for confirming or not confirming a hypothesis. Columns 2 to 5 list the incidence of considering the type-I-error and type-II-error.
Not confirmed: Hypothesis could not be confirmed by statistically significant test. Confirmed: Hypothesis or one side of the hypothesis could be confirmed by a statistically significant test. Both: both type-I-error and type-II-error were mentioned. Type-I-error: only the type-I-error was mentioned. Neither: neither the type-I-error, nor the type-II-error were mentioned. Type-II-error: only the type-II-error was mentioned

	Type-I-error	**Type-II-error**	**Both**	**Neither**	**Total**
Confirmed	7	0	43	5	55
Not confirmed	14	2	48	0	64
Total	21	2	91	5	119

Most of the studies considered both errors (76.5%, 91 of 119 studies).
Studies being able to confirm their hypothesis considered the type-I-error in 90.9% of the studies (50 out of 55 studies). The type-I-error is the risk of rejecting a true null hypothesis. Compared to that, studies, which could not confirm their hypothesis, considered the type-II-error in 78.1% of the studies (50 out of 64 studies). The type-II-error is the risk of not rejecting a false null hypothesis. All of the 5 studies, which did not comment on the risk of making a type-I-error or type-II-error, could confirm their hypothesis.

3.3.4 Other observations

We performed six different tests for concordance, but we abridged the results of the tests as we wanted to make the text readable. We collected a lot of data, but it did not increase the value of the work by comparing everything with everything else. In all comparisons with different variables, where an influence of more than one variable is expected, we abandoned the analysis as no reliable declarations can be made in any case.

This comparison focuses on the thorough depiction of the statistical methods by stating both the described error and the type of statistical testing by applying either one-sided or two-sided tests. We do not demonstrate the results in detail, but we did observe that studies applying a one-sided or two-sided test, considered both errors in 85.5% (65 out of 76 studies). Studies not stating the application of a one-or two-sided test did also forget the consideration of both the type-I-or type-II-error in 41.5% of the studies (17 out of 41).

The type of statistical test in studies was compared to the difference in the calculated and recruited sample size. The comparison of the two variables is relevant, as with a one-sided

testing a smaller sample size is required. With a two-sided test a bigger sample size is required as the confidence interval is 0.025 in two directions.
However, we do not expect any new information of the test of this concordance, because we know that there are only 8 studies with a one-sided test and six different options to file the 8 studies.

We investigated, if there is a relation of the confirmation of a hypothesis and the achievement of the calculated sample size. The detailed results will not be displayed as we figured a confirmation, or a missing confirmation of a hypothesis is influenced by more than one factor.
Only one observation can be stated. 9 out of 10 studies which did not recruit the calculated sample size were not able to confirm their hypothesis. One study could confirm the superiority of a combination of nutritional supplement over placebo (see study 43).

3.4 The difference between the selected journals

It was already commented on the diverging impact factors of the six different journals (see chapter 2.). We selected certain variables to compare the journals (see table 3.11).

Table 3.11: Results in comparing the six different journals

Explanation: The left column lists the selected criteria for comparing the journals. Column 2 to 7 shows the counted number for each journal, when the selected criteria was the case in one of the 20 or 19 papers. The papers are sorted by their descending number of the impact factor.
NEJM: New England Journal of Medicine. JAMA: Journal of the American Medical Association. Ann Surg: Annals of Surgery. Malaria: Malaria Journal. *: we did not include the studies in which the sample size was reached after a change of the protocol, because the change of a protocol requires a recalculation of the necessary sample size. **: studies which were early terminated according to protocol were included. ***: justified tests are a clearly stated superiority testing, a clearly stated justified non-inferiority and the testing of an equivalence. ****: unjustified tests have an unjustified use of a non-inferiority testing, not stated tests are studies without a clearly stated hypothesis.

	NEJM	JAMA	Ann Surg	Pediatrics	Malaria	Clinics
Impact Factor	59.6	37.7	8.6	5.5	3.1	1.3
Hypothesis clearly stated	15	17	13	16	15	7
One-sided test & unidirectional hypothesis	2	1	1	0	4	0
Two-sided test &bi-/multidirectional hypothesis	2	1	5	0	2	1
Number of recruited patients according to protocol * **	11	14	14	15	12	9
Type-I-& Type-II-error (Both)	16	18	17	17	11	12
Number of justified tests	15	17	13	16	14	7
Number of not justified tests and not stated tests	5	3	7	4	6	12
Proportion of confirmed study results in justified tests* **	10/15 66.7%	4/17 23.5%	7/13 53.8%	6/16 37.5%	9/14 64.3%	2/7 28.6%
Proportion of confirmed study results in not justified or not stated tests** **	5/5 100%	1/3 33.3%	2/7 28.6%	1/4 25%	4/6 66.7%	4/12 28.6%
Number of selected papers	20	20	20	20	20	19

We included journals with a rather high impact factor (NEJM, JAMA), journals being in medium range (Annals of Surgery, Pediatrics and the Malaria Journal) and one journal with a rather low impact factor (Clinics).

In almost all journals more than half of the studies clearly stated a hypothesis, except the studies from Clinics (7 out of 19 studies). It is the same with the number of studies which recruited patients according to protocol. Only Clinics did have less than half of the studies without the recruitment of the calculated sample size. NEJM and the Malaria Journal barely have more than half of the studies recruiting patients according to the protocol. Clinics also did have the highest number for unjustified tests and not stated tests. 12 out of 19 studies (63.2%) did not state their hypothesis or did use a non-inferiority testing where it was not applicable.

NEJM has the highest percentage of confirmed study results in not justified or not stated statistical tests. 5 out of 5 studies could confirm their study hypothesis, whereas 10 out of 15 studies with a justified test could confirm their study hypothesis in studies from NEJM. JAMA is the journal with the highest number of justified tests and in contrast the lowest proportion of confirmed study results in justified tests.

4 Discussion

Since the introduction of EBM, the importance of a clearly stated study question became apparent [39]. David Sackett and colleagues had the idea of a clearly stated hypothesis and intended to split it into four sections: patient, intervention, control and outcome.
A logical demand is, after a clearly stated question or hypothesis, that the objective and the tests of a study match the study question. In our small project the question of a study, its objective and tests were screened for concordance and a new catalogue of clinical trials was developed: the HOST catalogue. It includes the study hypothesis, objective, statistics and translation.

4.1 Summary of the main results

As our first question we investigated the description of a clear hypothesis in the selected studies. There were either unidirectional or bidirectional hypotheses expressed in the studies. 87.4% (104 out of 119) of the studies had a unidirectional and 12.6% (15 out of 119) a bidirectional or multidirectional hypothesis.
We also looked for the type of statistical testing in the studies. Either a one-sided testing (6.7%) or a two-sided (57.1%) testing was used. In total 78 studies stated their statistical test.

We then checked for concordances among the variables. 6.7% (7 out of 104 studies) of the studies with a unidirectional hypothesis used a one-sided statistical test. 66.7% (10 out of 15 studies) of the studies having a bidirectional hypothesis applied a two-sided statistical test.
Studies with a unidirectional hypothesis had a unidirectional study objective in 91.3% (95 out of 104), whereas studies with a bidirectional hypothesis had a bi- or multidirectional study objective in 73.3% (11 out of 15 studies).

5 out of 99 studies (5.1%) with a stated statistical test and a unidirectional study objective used a one-sided test. 9 out of 20 studies (45%) with a stated statistical test and a bi-or multidirectional study objective used a two-sided test.

12 out of the 78 studies (15.4%) with a stated statistical test had a concordant study hypothesis, study objective and statistical test.

4.2 Meaning of the findings and their importance

Many clinical investigators start with an asymmetric hypothesis. They assume that the experimental group is superior to the control group. In our project, 99 out of 119 studies had a unidirectional study objective for their tests. 95 of them formed the matching unidirectional hypothesis (96%). We counted 20 of 119 studies with a bi- or multidirectional study objective and 11 of them (55%) had the matching bidirectional study hypothesis. This forms a disproportionateness of 10.9% (13 of 119), where there should not be one. However, we can confirm that the majority of study objectives concord with the study hypothesis, merely the implementation of the statistical test is incorrect. Only 12 out of 78 (15.4%) studies, we excluded the studies which did not state a statistical test, applied the correct one- or two-sided test matching their uni- or bidirectional study hypothesis and study objective (see table 3.8). These results show a waste of statistical power, which could be easily diminished by applying the correct statistical test.
In general, we observed that the use of one-sided tests is very uncommon, although many trials we looked through demanded for a one-sided test. Only 8 out of 119 studies (6.7%) used a one-sided test and 99 out of 119 studies (83.2%) had a unidirectional study objective.

If you look at the literature handling the use of statistical tests, you seldom find papers that give a clear specification for the conditions to use a one-sided or two-sided test. One specification is from 1994 recommending a predominantly use of a two-sided test to cover both directions of possible results [3]. This is not justified from our point of view. The authors say it is necessary to actually cover and secure a result in both directions. This two-sided testing consumes a lot of statistical power which means a larger sample size in comparison to a one-sided testing [37]. The sample size can be reduced by requesting a clear distinction of a unidirectional hypothesis from a bidirectional hypothesis and to complete a one-sided test for a unidirectional hypothesis and a two-sided test for a bidirectional hypothesis. This will be a big step forward in the development of clinical trials.
“One sided tests should never be used simply as a device to make a conventionally non-significant difference significant.” [3, p.248] Blands and Altmans fear of getting more (false) significant results when using a one-sided test cannot be confirmed either. A one-sided test demands for half of the level of significance and therefore does not increase the risk of making a type-II-error [47].

Both authors say that a one-sided test is “justified by saying we are not interested in the possibility that the active treatment is worse than no treatment.” [3, p.248]. In our project this was often the case and 30 studies of 119 (25.2%) tested A vs. Placebo or A vs. ∅. Only one of these studies stated the use of a one-sided test.
Another argument for the two-sided testing of Bland and Altman is, “if a new treatment kills a lot of patients we should not simply abandon it; we should ask why this happened.” [3, p.248]. But why should we ask for it? What are we interested in, if it is not working? And should we increase the sample size only for this reason? In our opinion the answer is no. The knowledge about a treatment not being superior is a statement enough and should not lead to more patients being harmed or even killed. Studies with a hypothesis of superiority or non-inferiority are clearly testing in one direction and should therefore use a one-sided test. Studies testing the superiority of both treatments or are testing for the equivalence of the treatments should use a two-sided test. An example for the difficulty of finding the correct statistical testing or hypothesis is the comparison of a new with a standard treatment. Either we want to know if a new treatment is superior to the standard treatment or if the new treatment is equal to the standard treatment. The last example can be the case when the new treatment is cheaper than the standard and more expensive treatment. The authors therefore have to be aware of their hypothesis and aim of the study. Testing the superiority of the new treatment demands a one-sided test. Testing the equivalence of both treatments demands a two-sided test. And there is a third option: testing the non-inferiority of the new treatment again demands a one-sided test, as it is a unidirectional hypothesis.
In summary it can be said that most studies demand a for the use of a one-sided test. These are studies testing for superiority and non-inferiority of one of their treatments. Only studies which are testing the equivalence of treatments should use a two-sided test.
There is an exception when it comes to the use of two-sided tests. We observed that a few studies wanted to the test the superiorities for both their treatment arms. In this case we suggest using a two-sided test, which is theoretically a two-time performed one-sided test. However, the acceptance of two statistical tests again increases the number of patients that have to be included. As this question is frequently asked in many trials, we have to discuss that a superiority test alone may be sufficient. If the hypothesis of superiority will be rejected an additional test of non-inferiority has to be justified by any expected advantage or is not necessary at all.

We can even go further and combine our demand for the more frequent use of a one-sided test with the achievement of the calculated sample size. We counted a lot of studies with a unidirectional hypothesis and a two-sided test and therefore the requirement for the numbers of patients being recruited are higher than with a one-sided testing. If we limit ourselves to the use of one-sided tests in studies testing for a unidirectional hypothesis requiring only a one-sided test, we predict more studies achieving the calculated sample size. This can be said by looking at our results without the display of an extra table and should be discussed.
One example of a misconducted study is study 74 (see appendix), which appeared in Annals of Surgery. The authors wanted to prove the superiority of a neurectomy over a sham procedure (placebo). They used a two-sided α=0.05, although they clearly stated a unidirectional hypothesis. This is one example of inconsistencies in randomized controlled trials and requires more recruited patients than necessary.

Another observation is that the documentation of the statistical testing is not considered as important as it should be. 41 of 119 studies (34.5%) did not state their applied statistical test at all. Additionally, in 24 out of 119 studies (20.2%) the sample size was not documented. The calculated sample size serves as a parameter to detect possible problems in the realization of a study. It gives a hint how a study was conducted and in which way the results were reached and is therefore of high interest for the reader.
It has to be said that the documentation of the statistical testing lacks in the examined studies. In these cases, it is hard to get an impression of the statistical power and if the results are convincing. Almost the same situation can be seen by counting the studies not clearly stating their hypothesis. 30.3% (36 out of 119 studies) did no clearly state their hypothesis in words. A clearly stated hypothesis helps us to file the papers and to analyse the study objective and statistical methods.
All in all, it becomes clear that a complete and structured documentation is absolutely essential for the assessment of a study. Without a clearly stated hypothesis or without consistent variables, the translation of the study results is impaired, and the process of a study cannot be reenacted. We therefore plead for a clear statement of the generated hypothesis, the study objective and the statistical testing.
A good study documentation will give a good transparency for the reader and will improve clinical research sooner or later.

By looking at the calculated and recruited sample size again, there is another thing that needs to be discussed. 12 out of 119 studies (10%) changed their study protocol during the study implementation. 83.3% (10 out of 12) out of those achieved their calculated sample size. This can be seen as a mistake, because the required sample size was calculated on the basis of the original protocol. If the original protocol is changed, the study's baseline is deformed simultaneously. For example, if a change in the eligible study population results in 3 or more patients, the whole study situation is altered. This means a change of protocol entails a new power calculation and should be exercised this way in research. Again, a clearly described study objective and statistical test is essential for the conduct of a study. For studies that changed their protocol or terminated early we checked *cinicaltrial.gov*. It documents all the details which exceed the extent of a paper. We got the impression that *clinicaltrials.gov* is not always as reliable as it should be, as we were not always able to find the reasons for a change or an early termination of a study. 24 of 119 studies had a change of protocol or terminated early. In 10 of the 24 studies (41.7%) no reason was stated for an early termination or for not reaching the calculated sample size. *Clinicaltrials.gov* should provide the missing information. 7 out of these 10 RCTs (70%) with an early termination were not in line with the protocol found on *clinicaltrial.gov* and are listed as not reached due to early termination.

Another point of interest was the consideration of the type-I-error and type-II-error in the statistical analysis section. As most of the studies did consider both errors or at least one of them (95.8%, 114 of 119 studies), we claim that there is little risk of making a type-I- or type-II-error.
Literature often only features the risk of making a type-I-error [20] and at the same time emphasises the need for caring about the type-II-error [21]. Our study showed that there is only a small gap in the consideration of the type-I-error and type-II-error. 94.1% (112 of 119 studies) considered the type-I-error and 78.2% (93 of 119 studies) the type-II-error.

We hypothesized an incorrect study documentation is running the risk of getting biased results. It is not the same as a lack in documentation rather than the incorrect statement of the objective of the study.
12 studies had a concordant study objective, hypothesis and statistical test. 4 out of these 12 could confirm their hypothesis, whereas 8 could not confirm their hypothesis.

We also looked at the studies with a more elaborate design with more than two study arms and checked how many confirmed their study. Only 2 out of these 14 studies could confirm their hypothesis and, in both cases, they simultaneously checked a bidirectional superiority or superiority and non-inferiority. It was interesting to see that 12 out of 14 studies (85.7%) could not confirm their hypothesis.
In general, it is only a speculation, due to the small numbers, that a clearly stated study design lowers the chance to confirm the study hypothesis and biased results and therefore the risk of making a type-I-error. On the other side we can only speculate, due to the small numbers, that an unclear and rather chaotic study objective, such as A vs. B vs. AB vs. placebo (see appendix (study 101)), for example, also lowers the chance of confirming the study hypothesis.
A fact is that the numbers for confirming the hypothesis in these two examples is much lower than the total amount of the examined studies confirming their hypothesis: 46.2%.

In conclusion it cannot be said with our data if any inconsistencies lead to biased results.

4.3 The materials, methods and results in context of current research

The standard tool to evaluate the reporting quality of clinical studies by guidelines is the regularly updated CONSORT statement. We examined it to show which of its criteria was also investigated by us, which criteria we investigated in addition and if our criteria can be recommended as a supplement to the CONSORT statement.
The last CONSORT statement is from 2010 [41] and features a checklist of 25 items. It is a checklist to screen the correct and clear documentation of a study. An extension to the CONSORT-statement 2010 was published in 2012 [32] and takes a closer look on the reporting in noninferiority and equivalence trials. It was also included for comparison in the following chapter, as we also had RCTs with a hypothesis of noninferiority and equivalence.

Our finding that a clear hypothesis needs to be stated, is in line with the CONSORT-group. Item 2b in the subitem introduction of the CONSORT-statement demands the postulation of the specific objectives and hypotheses in the introduction.
The consideration of a clear documentation of changes to methods (see item 3b [41]) and outcomes after trial commencement (see item 6b [41]) corresponds partially with our investigation of the calculated sample size. We investigated if it could only be recruited after

a change in the study protocol. 10% of the studies we checked changed their study protocol after trial commencement. However, we did not investigate further details and reasons for any changes.

Subitems 4a and 13a focus on the criteria and number of participants. The CONSORT-group demands for specific information in the methods section on the eligibility criteria for participants and in the results section for the number of participants who were assigned, received the treatment and were analysed. In our project we only took account of the calculated sample size, if it was described and recruited, irrespective of the eligibility criteria and the further withdrawal of participants. However, the importance of just documenting the sample size is obvious: 20.2% of the studies did not describe their calculated sample size.

The CONSORT-statement also wants a clear documentation and explanation of any interim analyses and stopping guidelines and why a trial ended or was stopped (subitems 7b and 14b [41]). In our project we covered this specific item partially by having a closer look at the calculated sample size. 8.4% of the studies did not reach the sample size due to early termination or not stated reasons. 2.5% on the other hand terminated their study early according to protocol and given reasons.

By looking at the reported statistical methods, the CONSORT-statement of 2010 remains very superficial. The extension for noninferiority and equivalence trials is more specific [32]. It wants the information whether a 1- or 2-sided confidence interval approach was used, which we also documented in the 119 RCT's.

Another subitem (17a) [41, page 4] of the CONSORT-statement, which we also partially dealt with is the documentation of the primary and secondary outcomes, results for each group and the estimated effect size and its precision. In our project we handled it a lot more superficially and solely recorded if the hypothesis of a study could be confirmed or not.

A subitem which was also investigated by us, to some extent, is the documentation of limitations of a study. It is asked for the reporting of trial limitations, addressing sources of potential bias, imprecision and in some cases the multiplicity of analyses. We assessed the limitations of a study rather by the imprecisions in the concordance of hypothesis and its statistical testing. 39 of 119 trials used a two-sided test, although they clearly stated a superiority hypothesis. And to go even further: 57 of the 119 studies (48%) had a unidirectional study objective and used a two-sided testing.

The last subitem of the CONSORT-statement which we want to comment on is the interpretation of results (subitem 22). The extension for noninferiority and equivalence trials especially mentions the results should be interpreted in relation to the hypothesis. This again emphasizes the importance of a correctly reported hypothesis at the beginning of a trial and shows that the whole trial is built upon it.

Two of the six variables we looked at cannot be found in the checklist of the CONSORT-group. 15 of the 25 items the CONSORT-statement named were not documented by us. The CONSORT group laid special interest on the randomization and blinding process and details on the formal information about registration and funding. We, however, did not check those variables but looked for two others. First, we checked if the study objective has a uni- or bidirectional approach by looking at the comparators. In some sort the CONSORT-statement also wanted information on the comparators by seeking information on the intervention and how it was performed to be able to replicate the study (item 5). They, however, do not want any information on the intervention with regards to the study hypothesis and if it is in line with the study hypothesis.
We think item 5 should be revised as: “The interventions for each group with sufficient details to allow replication, including how and when they were actually administered” [41, p.3], taking into account the study hypothesis.
The second item we, but not the CONSORT-group, analysed is the consideration of the type-I-/-II-error. 76.5% (91 of 119) studies documented the testing of both errors. The great number is more than we expected, and we register it as a success in the documentation of clinical trials. Therefore, there is no need to especially mention the type-I-/-II-error in the statistical methods item number 12 of the CONSORT-statement.

Our aim to improve the quality of conducting an RCT is the same aim as of the CONSORT-group. They want to improve reporting of RCTs by pursuing the same objective as we do: clear information on the methodology and findings of trials to minimize biased results.
We have the impression that the question of a study or rather the hypothesis of a study is the key for the conduct of a study due to the CONSORT statement.
This matches with our demand for a clearly stated hypothesis.
To form a hypothesis a clinical question is needed for evaluation. A method to construct a clinical question is the theory of PICO [45] as stated in the introduction of the discussion.

PICO is an acronym for Patient/Problem, Intervention, Intervention for Comparison and Outcome. All of the four variables should be worked out to then form a study question. [39]
By taking a closer look, it became apparent that there is still need of a precise adjustment for a clearly stated study hypothesis. PICO is sufficient to formulate a study question, but it is not enough to form a study hypothesis. Possibilities and expectations need to be weighed out, so a clear direction of a study can emerge. The null and alternative hypothesis needs to be explicitly stated to also form the direction of the study objective and to be able to apply the correct statistical tests.
Our demand is the application of the HOST catalogue. It includes the four variables: study hypothesis, objective, statistical test and translation of the study results. It is the general framework for a study and prevents mistakes in the implementation and conduct of a trial by stating the concordant and matching parameters for a study.

A similar work published in 2005 by An-Wen Chan and Douglas Altman reports that most trials fail to specify their primary outcomes and this supports our findings [5]. They analyzed a lot more trials than we did (519 trials) and assessed the papers by using six different criteria. Two of them overlap with our acquisition. They assessed the power calculation. We assessed it by looking at the sample size. Our approach of counting the stated hypotheses can be compared with their specification of the primary outcome. Their conclusion highlights the need for improvement in the reporting of methodological details. We can confirm these needs of a clear documentation and execution of the study objective.

4.4 Alternative explanations

By searching for proof of the need of progress in Evidence Based Medicine it always has to be taken care of bias.
A source for mistakes is the journal Clinics we included. It is from Brazil and has the lowest impact factor (1.3) of the studies we included. It only represents 16% (19 out of 119 studies) of the examined studies but could have influenced the results we observed. 79% of the papers in Clinics could not confirm their hypothesis, whereas 75% of the NEJM papers were able to achieve significant results. Clinics is also the paper with the highest number of papers with a not clearly stated hypothesis, more than half of the studies did not state their hypothesis explicitly in words. And about one third of the studies in MJ (impact factor 3.1) and Clinics did not state their calculated sample size at all.

In comparison, almost all the papers in JAMA (impact factor 37.7) stated the use of a two-sided testing, whereas more than half of the studies in Clinics and half of the studies in Annals of Surgery (impact factor 8.6) did not state any form of a statistical test.
This can be one factor of letting the results we observed appear worse than the actual situation of more frequently quoted journals and can be counted as an alternative explanation and factor of bias.

The following request out of the findings above is that a published paper in a journal should at least state a small comment on the statistical analysis applied. In the end the journal's reader is not informed about essential information. A difference between the six journals in the requirements for publication can also be observed. Clinics and the Malaria Journal, the ones with least ranked impact factor, did accept more papers not mentioning both the type-I-error and type-II-error. It can be speculated that there is an additional concordance in the publication politics of the most to least cited journals and the accuracy in stating relevant information.

4.5 Clinical relevance

A clinical problem occurs and needs to be solved. More than 3500 papers were published on PubMed each day in 2017 to give solutions for clinical problems [8].
The right thing by conducting an RCT is to establish a clear study question and to evolve a study scheme by the PICO-idea and the HOST-catalogue. Every study needs a clear study question (objective and hypothesis) and the matching statistical test to prove or object the alternative hypothesis.
If authors of trials do not know which of the treatments tested is superior, one should be sufficiently realistic. In some cases, it is possible to demonstrate superiority or noninferiority of one of the treatments. In many cases it has to be expected that no difference is to be shown. In a situation where a superiority or noninferiority test was performed, and a confirmation could not be verified, the correct conclusion is we do not know! In this case, if preliminary evidence does not suggest which of the two methods will be superior, we recommend completing an equivalence test. This is always better than not clearly stating a hypothesis. 30.3% (36 of 119) of the trials did not clearly state a hypothesis. When a hypothesis is not clearly stated, it is almost impossible to decide, whether or not the authors considered the different statistical options.

Another important thing is the objective of a trial. When a clear study objective is phrased, the hypothesis of a trial needs to be carefully elaborated. A clearly stated objective should normally lead to a clearly stated hypothesis or the other way around. A clearly defined study hypothesis only gives the option for confirming the null hypothesis or the alternative hypothesis. It does not help, if a new treatment is compared to three or more different treatments or combinations of treatments. An example for not clearly sorted study objectives is an RCT with multiple comparators and study arms of more than 2. And as we expected only 3 out of 13 studies (23.1%) could confirm either the superiority or equivalence of one of their treatment arms. The rest were not able to confirm their hypothesis.

To draft a higher aim: every statistically significant data needs to pass in real life conditions. It needs proof of effectiveness and not only of efficacy. Efficacy means that the new product or tool really works under ideal study conditions. Efficacy is different from effectiveness. Effectiveness means the effects observed under ideal world conditions can be reproduced under real world conditions and solve a patient health problem. Effects that work under ideal study conditions – independent of the limitations of a statistical test – have to demonstrate effectiveness to quality as standards of medical care. This "gulf" [6, p.2], as Archibald L. Cochrane declared it already in 1972, should be considered carefully. Bradford Hill added in 1984: "At its best such a trial shows what can be accomplished with a medicine under careful observation and certain restricted conditions. The same results will not invariably or necessarily be observed when the medicine passes into general use;…" [19, p.3152]
To narrow this gap the pragmatic controlled trial is a way and a possibility to test the effectiveness under real world conditions [33]. The need for testing a treatment under real world conditions is indispensable and with this small experiment it was shown that there is lack of the expected and real implementation of studies or experiments.

4.6 Limitations & strengths

The question is if the 120 trials out of 6 journals represent the general state in research and if we could have chosen other more representative journals.
We selected six journals, which are being read regularly by our medical colleagues. The reason for this was to avoid an additional burden for our co-workers and to be able to perform our project. Since the co-workers had already read the editions of their journals, they

only had to document small details from the selected papers. We assumed that 20 selected papers for each co-worker would be manageable next to their daily workload.
We held the documentation of the six variables as simple as possible due to the fact that every result was read a second time by another co-worker.
It would, however, have been more representative with more collected trials, also considering that many studies did not state their statistical testing and could not be included in some evaluations. Furthermore, we cannot be sure if some subjective visions on the studies of the co-workers and myself, such as the derived hypothesis of studies, where none was clearly stated, could have biased the outcomes. Our expectation was that the documentation and concordance is lacking, so there was a preference to find information to support our hypothesis.

As the only variable the hypothesis of the study was derived out of the context. In case of the statistical test or the type of error it was concluded `not stated`. This could have biased the outcomes as mistakes can happen in the understanding of trials and misinterpretations of the study hypothesis. This could have led to false positive or false negative numbers with too many or less trails with a confirmed hypothesis.
It can be argued that the act of deriving the hypothesis was redundant and the study should have been excluded as `hypothesis not stated`. Due to the high numbers of trials with no statement of a hypothesis we needed the numbers of papers to get results.

At the same time, however, we have to say that a stated statistical test cannot be falsified. The results speak for themselves as we found 57 of 77 studies (74%) with a stated test not applying the matching statistical test to their hypothesis.
Our idea of using one-sided tests more frequently is new and has never been suggested before. Our argument of needing a smaller sample size was never an issue and should be paid attention to. Our project is the first one to investigate this misunderstanding and it provides solutions.

By searching for proof of the need of progress in Evidence Based Medicine it always has to be taken care of bias.
A source for mistakes is the journal Clinics we included. It is from Brazil and has the lowest impact factor (1.3) of the studies we included. It only represents 16% (19 out of 119 studies) of the examined studies but could have influenced the results we observed. 79% of

the papers in Clinics could not confirm their hypothesis, whereas 75% of the NEJM papers were able to achieve significant results. Clinics is also the paper with the highest number of papers with a not clearly stated hypothesis, more than half of the studies did not state their hypothesis explicitly in words. And about one third of the studies in MJ (impact factor 3.1) and Clinics did not state their calculated sample size at all.
In comparison, almost all the papers in JAMA (impact factor 37.7) stated the use of a two-sided testing, whereas more than half of the studies in Clinics and half of the studies in Annals of Surgery (impact factor 8.6) did not state any form of a statistical test.
This can be one factor of letting the results we observed appear worse than the actual situation of more frequently quoted journals and can be counted as a limitation and a factor of bias.
However, we can request out of the findings above is that a every published paper in a journal, no matter which impact factor it has, should at least state a small comment on the statistical analysis applied. In the end the journal's reader is not informed about essential information.

4.7 Suggestions for further research

Our project shows that a clearly statement on the study hypothesis is absolutely essential. Although it was the case in most of the studies, we recommend a clear and evident description of the study hypothesis. A further appeal is to adjust the statistical testing to the described study hypothesis. If a unidirectional hypothesis is investigated, a one-sided test must be applied.
This postulation is related to the quintessence of our project: We criticize that in the majority of the studies, when a one-sided test would have been legitimate due to a unidirectional study hypothesis and study objective, a two-sided test was applied. This implicates more patients need to be recruited than required, due to the fact that the implementation of a one-sided test requires less recruited patients than the implementation of a two-sided test.
A two-sided test should be applied when a bidirectional study hypothesis is drafted.

We therefore recommend the stringent adherence to the above-mentioned suggestions for planning and implementing clinical studies, to prevent the needless recruitment of more patients than required. This should be considered for ethical reasons and is also a cost factor.

Our proposal is controversial. Recommendations for or against one-sided tests are common and were common in the past [11]. Authors in the past have promoted the application of a two-sided testing [2].

John Martin Bland and Douglas G Altman published notes on statistics . We already said that they advised to use a two-sided test in 1994 and only use a one-sided test for good reasons, or if a difference has the same consequences as no difference [3]. They serve placebo-controlled trials as an example for a good reason. We suggest the same and furthermore claim to use one-sided tests in experiments comparing a treatment to no intervention (∅) or for experiments comparing a new treatment to an old and established standard treatment.

We furthermore hope to sharpen the minds of researchers for a clear documentation and a logical application of statistics.

Our recommendations should be controlled more thoroughly, and feedback needs to be given to the authors. Nevertheless, the publication process of not well documented trials is in the end in the hand of the editors of journals.

As one step prior to that the process of the selection and publication should be modified, and already medical students should become aware of the lack in medical research from the beginning and should be informed how they can prevent it. A tool for the future would be the pragmatic controlled trial. It supplies us with real life data on the effectiveness of treatments.

A clearly documented and conducted RCT is the first step, a pragmatic controlled trial is the second.

5 Summary

As an incidental finding of earlier scientific work, the impression was created that the described hypothesis in a clinical study was not always in concordance with the declared study objective and the statistical test applied. In addition, the assumption arose that the consideration of the type-I-error and the type-II-error was often not taken into account and that the calculated sample size was not met in clinical studies. Therefore, we set ourselves the task to describe the concordance of the study hypothesis with the study objective and with the application of the statistical tests in 120 studies from six different disciplines.
For this purpose, we identified 120 published randomized controlled trials with 20 publications each from six journals of different disciplines, beginning in November 2013 and going chronologically backwards.
Each study was classified on the basis of six criterions according to predefined classifications: First, it was determined whether the study hypothesis was a superiority, non-inferiority or equivalence hypothesis. Superiority and non-inferiority hypotheses are one-sided hypothesis tests, an equivalence test is a two-sided hypothesis test. Regarding the type of study objective, the comparison cohorts in the studies were used to decide whether a one-sided study objective or a two- or multi-sided study objective was pursued. For the type of statistical test, it was noted whether a one- or two-sided statistical test was applied in the studies. The fourth criterion we looked at was the consideration of the type-I-error and the type-II-error. Under the fifth and sixth criterion, the type of study result was classified on the basis of the comparison of the calculated and recruited sample and whether the study hypothesis was confirmed or not confirmed.
In a second step, the concordance of the one- or two-sided study hypothesis with the one-, two- or multi-sided study objective and the one- or two-sided statistical test was checked.

It was found that the concordance of the study hypothesis and the study objective was 89.1%. The concordance of the study hypothesis and the statistical test was 21.8% and the concordance of the study objective and statistical test was 17.9%.
In total, 15.4% of the studies had a concordant one- or two-sided study hypothesis with the matching one- or two-/multi-sided study objective and the matching one- or two-sided statistical test. Reviewing the studies showed that in 30.3% of the studies no clear statement on the study hypothesis could be found. 34.5% of the studies did not mention which type of statistical test they had used. The consideration of the type-I-error and type-II-error was

not described in 23.5% of the studies. 63% of the studies achieved the calculated sample size and 46.2% of the studies confirmed the study hypothesis.

Using the criteria collected from the 120 studies, we observed significant differences. Satisfactory was the high number of studies considering the type-I-error and type-II-error. It was equally pleasing that the majority of the studies were able to recruit the calculated sample size and almost half of the studies were able to confirm their study hypothesis. This is in contrast to the large number of studies that had an incomplete documentation of the examined criteria. We also found a deficit in the concordance of the study hypothesis with the study objective and with the applied statistical test.
Under ideal conditions, a complete concordance of the criteria would be expected. For example, the use of a one-sided test is stated when a one-sided superiority hypothesis is formulated, and the study is placebo-controlled with a one-sided study objective. A two-sided test should be applied if the study is testing an equivalence hypothesis with a two-sided study objective.
The most significant effect was seen in the considerable deviation in studies using a two-sided test and simultaneously formulating a one-sided hypothesis. This is explained by the fact that in 1994 statisticians recommended the predominant use of two-sided tests. This, however, means that more patients are included in an experimental study and more patients are exposed to a human experiment. That is the reason why we are discussing if we should harmonise these three criteria. It would lead to savings in patient numbers, thus minimising the risks for patients and shortening the duration of studies. In addition, the costs of conducting a study are reduced.

We have recorded the results of our project under the so-called HOST catalogue: we plead for the clear statement of the study Hypothesis and the study Objective, the matching application of the Statistical testing and the correct Translation of the study results.
In order to make studies more ethical, we propose to ensure that the clearly stated study hypothesis is in concordance with the clearly stated study objective and in concordance with the statistical test when conducting a study.

6 Bibliography

1. Altman DG: The scandal of poor medical research. British Medical Journal 308: 283-284 (1994)

2. Armitage P, Berry G, Matthews JNS: Statistical Methods in Medical Research, Fourth Edition. Blackwell Publishing, Oxford, 89 (2002)

3. Bland JM, Altman DG: Statistics Note, One and two sided tests of significance. British Medical Journal 309: 248 (1994)

4. Bradford Hill A: Principles of medical statistics - The aim of the statistical method. The Lancet Jan. 2: 41-43 (1937)

5. Chan AW, Altman DG: Epidemiology and reporting of randomized trials published in PubMed journals. Lancet 365: 1159-1162 (2005)

6. Cochrane AL: Effectiveness and Efficiency – Random Reflections On Health Services. The Nuffield Provincial Hospitals Trust, 2 (1972)

7. Concato J, Shah N, Horwitz R: Randomized, Controlled Trials, Observational Studies, and the Hierarchy of Research Designs. New England Journal of Medicine 342: 1887-1892 (2000)

8. Corlan DA: http://dan.corlan.net/medline-trend.html (19.12.2018)

9. Crocetti MT, Amin DD, Scherer R: Assessment of Risk of Bias Among Pediatric Randomized Controlled Trials. Pediatrics 126: 298- 305 (2010)

10. Davidoff F, Haynes B, Sackett D, Smith R: Evidence Based Medicine. British Medical Journal 310: 1085-1086 (1995)

11. Dubey SD: Some thoughts on the one-sided and two-sided tests. Journal of Biopharmaceutical Statistics 1: 139-150 (1991)

12. Fisher RA: The Principles of Experimentation, illustrated by a psycho-physical Experiment; Fisher RA: The Design of Experiments. Hafner Publishing Company, New York, 11-26 (1935)

13. Flacco ME, Manzoli L, Boccia S, Capasso L, Alekskova K, Rosso A, Scaioli G, De Vito C, Siliquini R, Villari P, Ioannidis JPA: Head-to-head randomized trials are mostly industry sponsored and almost always favor the industry sponsor. Journal of Clinical Epidemiology 68: 811-820 (2015)

14. Goldacre B, Heneghan C: How medicine is broken, and how we can fix it. British Medical Journal 350: h3397 (2015)

15. Greenhalgh T, Howick J, Maskrey N: Evidence Based Medicine: a movement in crisis? British Medical Journal 348: g3725 (2014)

16. Gugiu PC, Gugiu MR: A Critical Appraisal of Standard Guidelines for Grading Levels of Evidence. Evaluation& the Health Professions 33: 233- 255 (2010)

17. Guyatt GH, Oxman AD, Vist G, Kunz R, Brozek J, Alonso-Coello P, Montori V, Akl EA, Djulbegovic B, Falck-Ytter Y, Norris SL, Williams Jr. JW, Atkins D, Meerpohl J, Schünemann H: GRADE guidelines: 4. Rating the quality of evidence- study limitations (risk of bias). Journal of Clinical Epidemiology 64: 407-415 (2011)

18. Heneghan CJ: https://soundcloud.com/carl-heneghan/better-evidence-for-better-healthcare (18.05.2017)

19. Horton R: Common sense and figures: the rhetoric of validity in medicine. Bradford Hill memorial lecture 1999. Statistics in Medicine 19: 3149-3164 (2000)

20. ICH Expert Working Group: ICH Harmonised Tripartite Guideline, Structure and Content of Clinical Study Reports E3. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, step 4 version (1995)

21. ICH Expert Working Group: ICH Harmonised Tripartite Guideline, Statistical Principles for Clinical Trials E9. International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use, step 4 version (1998)

22. Ioannidis JPA: Effect of the Statistical Significance of Results on the Time to Completion and Publication of Randomized Efficacy Trials. Journal of American Medical Association 279: 281- 286 (1998)

23. Ioannidis JPA: Why most published research findings are false. PLOS Med 2 (8): e124 (2005)

24. Jefferson T, Jones M, Doshi P, Spencer EA, Onakpoya I, Heneghan CJ: Oseltamivir for influenza in adults and children: systematic review of clinical study reports and summary of regulatory comments. British Medical Journal 348: g2545 (2014)

25. Journal of American Medical Association: http://jamanetwork.com/journals/jama/pages/for-authors (23.05.2017)

26. MacCarthy A, Kirtley S, De Beyer JA, Altman DG, Simera I: Reporting guidelines for oncology research: helping to maximize the impact of your research. British Journal of Cancer 118: 619, 619-628 (2018)

27. Marshall G, Blacklock JWS, Cameron C, Capon N, Cruickshank R, Gaddum JH, Heaf FRG, Bradford Hill A, Houghton LE, Clifford Hoyle J, Raistrick H, Scadding JG, Tytler WH, Wilson GS, D'Arcy Hart P: Streptomycin treatment of pulmonary tuberculosis: A Medical Research Council investigation. British Medical Journal 2 (4582): 769-782 (1948)

28. Malaria Journal: https://malariajournal.biomedcentral.com/about (23.05.2017)

29. Meinecke AK, Welsing P, Kafatos G, Burke D, Trelle S, Kubin M, Nachbaur G, Egger M, Zuidgeest M: Series: Pragmatic trials and real world evidence: Paper 8. Data collection and management. Journal of Clinical Epidemiology 91: 13-22 (2017)

30. New England Journal of Medicine: http://www.nejm.org/page/media-center/fact-sheet (23.05.2017)

31. Ovid: http://www.ovid.com/site/catalog/journals/608.jsp#horizontalTab2 (23.05.2017)

32. Piaggio G, Elbourne DR, Pocock SJ, Evans SJW, Altman DG: Reporting of Noninferiority and Equivalence Randomized Trials, Extension of the CONSORT 2010 Statement. Journal of American Medical Association 308: 2594-2604 (2012)

33. Porzsolt F, Galito Rocha N, Toledo-Arruda AC, Thomaz TG, Moraes C, Bessa-Guerra TR, Leão M, Migowski A, Araujo da Silva AR, Weiss C: Efficacy and effectiveness trials have different goals, use different tools, and generate different messages. Pragmatic and Observational Research 6: 47-54 (2015)

34. Porzsolt F, Kliemt H: Ethische und empirische Grenzen randomisierter kontrollierter Studien. Medizin Klinik 103: 836-842 (2008)

35. Pub Neuro: http://www.pubneuro.com/brazilian-neurosurgery (23.05.2017)

36. Rosenberg W, Donald A: Evidence based medicine: an approach to clinical problem-solving. British Medical Journal 310: 1122-1126 (1995)

37. Ryan TP: Sample Size Determination and Power. Wiley Series in Probability and Statistics. Hoboken: John Wiley& Sons, Inc., Hoboken, New Jersey, p. 92 (2013)

38. Sackett DL: Evidence-Based Medicine. Seminars in Perinatology 21: 3-5 (1997)

39. Sackett DL, Richardson WS, Rosenberg W, Haynes RB: Evidence-based Medicine: How to Practice and Teach EBM. Churchill Livingstone, New York Edinburgh London Madrid Melbourne San Francisco Tokyo, 22-30 (1997)

40. Schulz KF, Chalmers I, Hayes RJ, Altman DG: Empirical Evidence of Bias Dimensions of Methodological Quality Associated With Estimates of Treatment Effects in Controlled Trials. Journal of American Medical Association 273: 408-412 (1995)

41. Schulz KF, Altman DG, Moher D: CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. British Medical Journal 340: c332 (2010)

42. Smith DG, Clemens J, Crede W, Harvey M, Gracely EJ: Impact of multiple comparisons in randomised clinical trials. The American Journal of Medicine 83: 545- 550 (1987)

43. Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al.: Dissemination and publication of research findings: an updated review of related biases. Health Technol Assess 14: (2010)

44. Stelfox HT, Chua G, O'Rourke K, Detsky AS: Conflict of Interest in the Debate over Calcium-Channel Antagonists. The New England Journal of Medicine 338: 101-106 (1998)

45. Straus SE, Sackett DL: Getting research findings into practice. Using research findings in clinical practice. British Medical Journal 317: 339-342 (1998)

46. Wang A, McCoy C, Murad MH, Montori VM: Association between industry affiliation and position on cardiovascular risk with rosiglitazone: cross sectional systematic review. British Medical Journal 340: c1344 (2010)

47. Weiß C: (23rd May 2017 personal notification per mail)

48. Wikipedia: https://en.wikipedia.org/wiki/Pediatrics_(journal) (23.05.2017)

Appendix

Table 6.1: Results for each examined study sorted by journal (first and last read) and date of appearance (new to old).

Journal; Year; Issue: First page – last page	Type of Comparator	Type of Hypothesis	Type of Statistical Test	Type of Statistical Confirmation	Type of described error	The difference in the calculated and recruited sample size	Annotations
1. NEJM; 2013; 369: 1115-2	A vs. AB New vs. Old	Superiority	Not stated	Superiority confirmed	Both	Early terminated according to protocol	Recruitment 2008-2013
2. NEJM; 2013; 369: 1124-33	A vs. B	Not stated {Superiority for A or B, bidirectional}	Two-sided	Not stated {Superiority confirmed}	Both	Reached as planned	Questionnaire Recruitment 2008-2010
3. NEJM; 2013; 369: 999-1010	AB vs. A+ Placebo New vs. Old	Superiority	Two-sided	Superiority not confirmed	Both	Not reached due to early termination	Recruitment 2009-2012
4. NEJM; 2013; 369: 799-808	A vs. B New vs. Old	Non-Inferiority (justified)	One-sided Alpha-level of 0.025	Non-Inferiority confirmed	Both	Reached after change of protocol	Δ-margin mentioned Recruitment 2008-2012
5. NEJM; 2013; 369: 438-47	A. vs. ∅	Superiority	One- & Two-sided Two-sided overall survival One-sided for sample size calculation	Superiority confirmed	Both	Reached after change of protocol	∅=observation Recruitment 2007-2010
6. NEJM; 2013; 369: 417-27	A vs. B New vs. Old	Non-inferiority & Superiority (Non-Inferiority justified)	Not stated	Non-Inferiority confirmed Superiority not confirmed	Type-I-error	Sample size not described	Δ-margin mentioned Recruitment 2004-2008
7. NEJM; 2013; 369: 319-29	A vs. Placebo	Not stated {Superiority}	Two-sided	Not stated {Superiority confirmed}	Both	Reached as planned	Recruitment 2009-2012

8. NEJM; 2013; 368: 2366- 76	A vs. ∅	Superiority	Not stated	Superiority not confirmed	Type-II-error	Reached as planned	∅=watchful waiting Recruitment 2008-2011
9. NEJM; 2013; 369: 213-23	A vs. Placebo	Not stated {Superiority}	Two-sided	Not stated {Superiority confirmed}	Both	Early terminated according to protocol	Recruitment 2008-2011
10. NEJM; 2013; 369: 111-21	AB vs. AC New vs. Old	Non-Inferiority (justified)	Two-sided	Non-Inferiority confirmed	Both	Reached as planned	Recruitment 2007-2010
11. NEJM; 2013; 368: 2084-93	A vs. B New vs. Old	Superiority	Two-sided	Superiority confirmed	Both	Early terminated according to protocol	Clinical trial gov: states it has been terminated after 2nd prespecified interim analysis Recruitment 2009-2013
12. NEJM; 2013; 369: 1227-36	A vs. Placebo	Not stated {Superiority}	Not stated	Not stated {Superiority confirmed}	Neither	Reached after change of protocol	Recruitment 2009-2011
13. NEJM; 2013; 369: 1317-26	A vs. Placebo	Non-Inferiority & Superiority (Non-Inferiority justified)	Not stated	Non-Inferiority confirmed Superiority not confirmed	Type-I-error	Sample size not described	{nonsense study, concerning the design} Recruitment 2010-2011
14. NEJM; 2013; 369: 1295-305	A vs. B	Superiority for A or B	Two-sided	Superiority not confirmed	Both	Reached as planned	Recruitment 2003-2011
15. NEJM; 2013; 369: 1395-405	A vs. ∅	Superiority	Two-sided	Superiority not confirmed	Both	Not reached due to early termination	∅= CRT turned off Recruitment 2008-2013
16. NEJM; 2013; 369: 1406-15	A vs. B New vs. Old	Non-Inferiority (justified)	Two-sided	Non-Inferiority confirmed	Both	Reached as planned	Recruitment 2010-2012
17. NEJM; 2013; 369: 1522-8	AB vs. A+ Placebo New vs. Old	Superiority	Two-sided	Superiority confirmed	Both	Reached as planned	Recruitment 2005-2010
18. NEJM; 2013; 369: 1491-501	A vs. B	Non-Inferiority & Superiority (Non-Inferiority justified)	One-sided for Non-Inferiority & Two-sided for Superiority One-sided: p-value of 0.025	Non-Inferiority confirmed Superiority not confirmed	Both	Reached as planned	Non-Inferiority (for mortality) Superiority (for COPD exacerbation) Problem of two primary endpoints Recruitment 2010-2011

19. NEJM; 2013; 369: 1587-97	AB vs. A New vs. Old	Superiority	Two-sided	Superiority not confirmed	Both	Reached after change of protocol	Recruitment 2008-2009
20. NEJM; 2013; 369: 1691-703	AB vs. A New vs. Old	Not stated {Superiority}	Two-sided	Not stated {Superiority confirmed}	Both	Reached after change of protocol	Recruitment 2009-2012
21. Malaria Journal; 2013; 12: 254	A vs. B	Not stated {Equivalence}	Two-sided	Not stated {Equivalence confirmed}	Type-I-error	Reached after change of protocol	Recruitment 2010-2011
22. Malaria Journal; 2013; 12: 55	A vs. B	Equivalence	Not stated	Equivalence not confirmed	Type-I-error	Sample size not described	Superiority design instead of equivalence design (Did not conduct an equivalence-study, no margins stated) It was planned to demonstrate equivalence but they used a design to investigate superiority. Recruitment 2010-2011
23. Malaria Journal; 2013; 12: 81	A vs. ∅	Superiority	Two-sided	Superiority confirmed	Both	Reached as planned	∅=no intervention Survey used for outcome measure Recruitment 2008-2009
24. Malaria Journal; 2013; 12: 251	A vs. B	Non-Inferiority (justified)	One-sided 5% significance level	Non-Inferiority confirmed	Both	Reached as planned	Recruitment 2008-2009
25. Malaria Journal; 2013; 12: 102	A vs. ∅	Not stated {Superiority}	Not stated	Not stated {Superiority confirmed}	Neither	Sample size not described	∅=obtained no ITNs ITNs in communities that partially already had ITNs Recruitment 2010-2011
26. Malaria Journal; 2013; 12: 363	A vs. ∅	Superiority	Not stated	Superiority not confirmed	Both	Reached as planned	∅= no ITNs Recruitment 1997-1998
27. Malaria Journal; 2013; 12: 79	AB vs. A	Not stated {Superiority}	One-sided Significance level of 0.05	Not stated {Superiority not confirmed}	Type-I-error	Reached as planned	Recruitment not stated
28. Malaria Journal; 2012; 11: 8	AB vs. A New vs. Old	Superiority	Not stated	Superiority confirmed	Both	Reached as planned	Recruitment 2009-2010

29. Malaria Journal; 2012; 11: 73	A vs. ∅	Not stated {Superiority}	Two-sided	Not stated {Superiority confirmed}	Both	Reached after change of protocol	∅=no preventive treatment Recruitment 2006-2008
30. Malaria Journal; 2012; 11: 174	A vs. B New vs. Old	Non-Inferiority (justified)	One-sided Significance level of 5%	Non-Inferiority confirmed	Both	Reached as planned	Recruitment 2008-2009
31. Malaria Journal; 2012; 11: 433	A vs. B New vs. Old	Non-Inferiority (justified)	Two-sided	Non-Inferiority confirmed	Both	Reached as planned	Recruitment 2008-2009
32. Malaria Journal; 2012; 11: 364	A vs. B New vs. Old	Non-Inferiority (justified)	Two-sided	Non-Inferiority confirmed	Both	Reached as planned	Recruitment 2007-2008
33. Malaria Journal; 2012; 11: 150	A vs. B	Non-Inferiority (justified)	Not stated	Non-Inferiority confirmed	Neither	Sample size not described	Recruitment 2006-2009
34. Malaria Journal; 2012; 11: 416	A vs. B New vs. Old	Non-Inferiority (justified)	Two-sided	Non-Inferiority confirmed	Both	Reached as planned	Recruitment 2010
35. Malaria Journal; 2011; 10: 148	AB vs. A+ Placebo New vs. Old	Superiority	Not stated	Superiority not confirmed	Type-I-error	Sample size not described	Recruitment 2004-2006
36. Malaria Journal; 2011; 10: 50	A vs. B New vs. Old	Non-Inferiority (not justified)	Two-sided	Non-Inferiority confirmed	Type-I-error	Sample size not described	No comment on the clinical justification of DHA/PQP to perform a non-inferiority trial. Recruitment 2005-2006
37. Malaria Journal; 2011; 10: 237	A vs. B	Non-Inferiority (justified)	One-sided 2.5% significance level	Non-Inferiority confirmed	Type-I-error	Reached as planned	Recruitment 2007-2009
38. Malaria Journal; 2011; 10: 387	A vs. ∅	Superiority	Not stated	Superiority not confirmed	Both	Reached as planned	∅=no preventive treatment Recruitment 2005-2007
39. Malaria Journal; 2011; 10: 247	A vs. Placebo	Superiority	Not stated	Superiority not confirmed	Both	Reached as planned	Recruitment 2007-2008
40. Malaria Journal; 2011; 10: 231	A vs. B vs. C	Not stated {Superiority for A, B or C}	Two-sided	Not stated {Superiority not confirmed}	Type-I-error	Sample size not described	Recruitment 2005

41. Clinics; 2013; 68 (11): 1400-1407	A vs. B New vs. Old	Superior-ity	Two-sided	Superior-ity not confirmed	Both	Reached as planned	
42. Clinics; 2013; 68(10): 1318-1324	AB vs. AC vs. AD	Not stated {Equiva-lence}	One-sided Signifi-cance level of 5%	Not stated {Equiva-lence con-firmed}	Both	Reached as planned	
43. Clinics; 2013; 68(3): 351-358	A vs. B vs. C vs. AB vs. AC vs. BC vs. ABC vs. Pla-cebo	Not stated {Superior-ity for groups over Pla-cebo}	Two-sided	Not stated {Superior-ity con-firmed}	Both	Not reached due to un-known rea-sons	Underpowered
44. Clinics; 2013; 68(3): 323-328	A vs. B New vs. Old	Superior-ity	Two-sided	Superior-ity con-firmed	Type-I-error	Sample size not described	
45. Clinics; 2012; 67(9): 1035-1038	A vs. B New vs. Old	Non-Infe-riority (justified)	Not stated	Non-Infe-riority confirmed	Both	Reached as planned	Superiority of the new cutter was stated in the conclusion section in terms of anastomotic bleeding.
46. Clinics; 2012; 67(9): 1059-1062	A vs. B	Not stated {Superior-ity for A or B}	Not stated	Not stated {Superior-ity not con-firmed}	Type-I-error	Sample size not described	
47. Clinics; 2012; 67(10): 1149-1155	AA vs. AB vs. BC New vs. Old	Superior-ity	Two-sided	Superior-ity not confirmed	Both	Reached as planned	
48. Clinics; 2012; 67(12): 1407-1414	A vs. BB vs. BC New vs. Old	Not stated {Superior-ity}	Two-sided	Not stated {Superior-ity not con-firmed}	Both	Not reached due to un-known rea-sons	Underpowered
49. Clinics; 2011; 66(8): 1353-1360	KEIN RCT						
50. Clinics; 2011; 66(8): 1321-1327	A vs. ∅	Not stated {Superior-ity}	Not stated	Not stated {Superior-ity not con-firmed}	Type-I-error	Sample size not described & Change of protocol	∅=no treatment
51. Clinics;	A vs. B vs. C	Not stated	Not stated	Not stated	Both	Reached as planned	

2011; 66(5): 811-815		{Superiority for A, B or C}		{Superiority not confirmed}			
52. Clinics; 2012; 67(9): 1029-1034	A vs. B vs. C New vs. Old	Superiority	Not stated	Superiority not confirmed	Type-I-error	Sample size not described	
53. Clinics; 2012; 67(8): 871-875	A vs. B New vs. Old	Not stated {Superiority for A or B}	Not stated	Not stated {Superiority not confirmed}	Both	Reached as planned	Questionnaire
54. Clinics; 2012; 67(5): 469-474	A vs. B New vs. Old	Not stated {Superiority}	Not stated	Not stated {Superiority confirmed}	Type-I-error	Sample size not described	Treatment B was evaluated to function as Placebo.
55. Clinics; 2012; 67(1): 49-54	A vs. B	Superiority	Not stated	Superiority not confirmed.	Both	Reached as planned	
56. Clinics; 2012; 67(1): 11-18	A vs. B vs. AB vs. AC vs. BC vs. ABC	Not stated {Superiority}	Not stated	Not stated {Superiority not confirmed}	Type-I-error	Sample size not described	The aim of the study was to evaluate the possible benefit of genetics (Apolipoprotein 4) not necessarily the difference between the study groups.
57. Clinics; 2011; 66(6): 1003 – 1007	A vs. B New vs. Old	Not stated {Superiority}	Not stated	Not stated {Superiority not confirmed}	Type-I-error	Sample size not described	
58. Clinics; 2011; 66(7): 1187-1191	A vs. Placebo	Not stated {Superiority}	Two-sided	Not stated {Superiority confirmed}	Both	Sample size not described	A saline solution functioned as Placebo.
59. Clinics; 2011; 66(12): 2001-2005	A vs. B New vs. Old	Not stated {Superiority}	Not stated	Not stated {Superiority not confirmed}	Both	Reached as planned	
60. Clinics; 2011; 66(10): 1721-1727	AB vs. A New vs. Old	Superiority	Not stated	Superiority not confirmed	Both	Reached as planned	Hypothesis not clearly stated, unclear when primary outcome defines a result of superiority or inferiority (in this trial: Superiority of AB concerning inspiratory muscle strength)

61. Annals of Surgery; 2013; 258: 690- 695	A vs. B New vs. Old	Not stated {Superiority for A or B}	Two-sided	Not stated {Superiority not confirmed}	Both	Reached as planned	
62. Annals of Surgery; 2013; 258: 527-533	A vs. B	Superiority	Two-sided	Superiority not confirmed	Both	Not reached due to unknown reasons & Change of protocol	
63. Annals of Surgery; 2013; 258: 385-393	A vs. B	Equivalence	Two-sided	Equivalence not confirmed	Both	Reached as planned	
64. Annals of Surgery; 2013; 257: 1016-1024	A vs. ∅	Superiority	Not stated	Superiority not confirmed	Both	Reached after change of protocol	∅=no sham feeding with gum
65. Annals of Surgery; 2013; 257:1025-1031	AB vs. A New vs. Old	Superiority	Not stated	Superiority confirmed	Neither	Sample size not described	
66. Annals of Surgery; 2013; 258: 21-29	A vs. ∅	Not stated {Superiority for A or ∅}	Two-sided	Not stated {Superiority not confirmed}	Both	Reached as planned	∅=no T-tube
67. Annals of Surgery; 2013; 258: 30-36	A vs. ∅ New vs. Old	Not stated {Superiority}	Two-sided	Not stated {Superiority not confirmed}	Type-I-error	Reached as planned	Complicated hypothesis: hard to imagine that it was defined prospectively. ∅=standard care
68. Annals of Surgery; 2013; 258: 37-45	A vs. B	Not stated {Superiority for A or B}	Two-sided	Not stated {Superiority not confirmed}	Both	Reached as planned	
69. Annals of Surgery; 2013; 258: 46-52	A vs. B vs. AB	Superiority	Not stated	Superiority not confirmed	Both	Reached as planned	
70. Annals of Surgery; 2013; 258: 53-58	A vs. B New vs. Old	Not stated {Equivalence}	Two-sided	Not stated {Equivalence confirmed}	Both	Reached as planned	
71. Annals of Surgery; 2013; 258: 240-247	A vs. B New vs. Old	Superiority	Not stated	Superiority confirmed	Both	Reached as planned	

72. Annals of Surgery; 2013; 258: 248-256	A vs. B New vs. Old	Superiority	One-sided Significance level of α=0.05	Superiority not confirmed	Both	Reached after change of protocol	
73. Annals of Surgery; 2013; 257: 834-838	A vs. ∅	Superiority	Not stated	Superiority confirmed	Neither	Sample size not described	∅=no music
74. Annals of Surgery; 2013; 257: 845-849	A vs. Placebo	Superiority	Two-sided	Superiority confirmed	Both	Reached as planned	Placebo=sham operation
75. Annals of Surgery; 2013; 257: 938-942	A vs. B	Superiority	Not stated	Superiority confirmed	Both	Reached as planned	
76. Annals of Surgery; 2013; 257: 839-844	A vs. ∅	Superiority	Two-sided	Superiority confirmed	Both	Reached as planned	∅=no feedback
77. Annals of Surgery; 2013; 257: 390-399	A vs. B New vs. Old	Superiority	Two-sided	Superiority confirmed	Both	Reached as planned	
78. Annals of Surgery; 2013; 257: 413-418	A vs. B New vs. Old	Not stated {Superiority}	Not stated	Not stated {Superiority not confirmed}	Both	Sample size not described	
79. Annals of Surgery; 2013; 257: 419-426	A vs. B New vs. Old	Not stated {Superiority}	Not stated	Not stated {Superiority confirmed}	Both	Reached as planned	
80. Annals of Surgery; 2013; 257: 214-218	A vs. B New vs. Old	Superiority	Not stated	Superiority not confirmed	Both	Reached as planned	
81. Pediatrics; 2013; 132 (5): e1163-e1172	A vs. B New vs. Old	Superiority	Two-sided	Superiority not confirmed	Both	Reached as planned	
82. Pediatrics; 2013; 132 (5); e1247-e1256	A vs. AB New vs. Old	Superiority	Two-sided	Superiority not confirmed	Both	Reached as planned	
83. Pediatrics; 2013; 132 (4): e832-e840	A vs. Placebo	Not stated {Superiority}	Not stated	Not stated {Superiority not confirmed}	Both	Sample size not described	

84. Pediatrics; 2013; 132 (4): e886-e894	A vs. B New vs. Old	Superior-ity	Two-sided	Superior-ity not confirmed	Type-I-error	Sample size not described	Old=active comparison group
85. Pediatrics; 2013; 132 (4): e895-e904	A vs. B New vs. Old	Superior-ity	Not stated	Superior-ity not confirmed	Both	Reached as planned	B=usual care
86. Pediatrics; 2013; 132 (4): e932-e938	AB vs. A + Placebo New vs. Old	Not stated {Superior-ity}	Two-sided	Not stated {Superior-ity not con-firmed}	Both	Reached as planned	
87. Pediatrics; 2013; 132 (5): e1236-e1245	A vs. B New vs. Old	Not stated {Superior-ity}	Not stated	Not stated {Superior-ity con-firmed}	Both	Reached as planned	B=usual care
88. Pediatrics; 2013; 132 (3): e623-e629	A vs. AB New vs. Old	Superior-ity	Not stated	Superior-ity con-firmed	Both	Reached as planned	
89. Pediatrics; 2013; 132 (3) e656-e661	A vs. AB New vs. Old	Superior-ity	Two-sided	Superior-ity con-firmed	Both	Reached as planned	
90. Pediatrics; 2013; 132 (3): e695-e703	A vs. Placebo	Superior-ity	Two-sided	Superior-ity not confirmed	Type-I-error	Sample size not described	
91. Pediatrics; 2013; 132 (4): e810-e816	AB vs. A + Placebo New vs. Old	Superior-ity	Two-sided	Superior-ity con-firmed	Both	Reached as planned	
92. Pediatrics; 2013; 132 (2): 326-331	A vs. ∅	Superior-ity	Not stated	Superior-ity not confirmed	Type-II-error	Reached as planned	∅=no insert
93. Pediatrics; 2013; 132 (2): e381-e388	A vs. B New vs. Old	Superior-ity	Two-sided	Superior-ity not confirmed	Both	Not reached due to early termination	Trialregister.nl states that the trial has been stopped because of slow recruit-ment and futility
94. Pediatrics; 2013; 132 (2): e389-e395	A vs. B New vs. Old	Superior-ity	Two-sided	Superior-ity not confirmed	Both	Reached as planned	
95. Pediatrics; 2013; 132 (1): e46- e52	A vs. Placebo	Superior-ity	Not stated	Superior-ity con-firmed	Both	Reached as planned	
96.	A vs. B New vs. Old	Superior-ity	Two-sided	Superior-ity not confirmed	Both	Reached as planned	Hypothesis not clearly stated, unclear when pri-mary outcome defines a

Pediatrics; 2013; 132 (1): e109-e118							result of superiority or inferiority; Old= usual care
97. Pediatrics; 2013; 132 (1): e119-e127	A vs. B vs. Placebo New vs. Old	Superiority	Two-sided	Superiority confirmed	Both	Reached as planned	
98. Pediatrics; 2013; 132 (1): e128-e134	A vs. AB New vs. Old	Superiority	Two-sided	Superiority confirmed	Both	Reached as planned	
99. Pediatrics; 2013; 132 (1): e135-e141	A vs. AB New vs. Old	Not stated {Non-Inferiority (not justified)}	Two-sided	Not stated {Non-Inferiority not confirmed}	Both	Not reached due to early termination	No comment on controlled-trials.com on the early study termination.
100. Pediatrics; 2013; 132 (1): e158-e166	A vs. B	Superiority	Not stated	Superiority not confirmed	Both	Reached as planned	
101. JAMA; 2013; 310(20): 2154-2163	A vs. B vs. AB vs. Placebo New vs. Old	Superiority	Two-sided	Superiority not confirmed	Both	Reached as planned	
102. JAMA; 2013; 310(20): 2164-2173	A vs. Placebo	Superiority	Two-sided	Superiority confirmed	Both	Reached as planned	
103. JAMA; 2013; 310(20): 2174-2183	A vs. B New vs Old	Superiority	Two-sided	Superiority not confirmed	Both	Not reached due to early termination	Clinicaltrials.gov: sponsor stopped the study for the security problem Old=standard care
104. JAMA; 2013; 310(19): 2050-2060	A vs. B New vs Old	Superiority	Two-sided	Superiority confirmed	Both	Reached as planned	Old=control group
105. JAMA; 2013; 310(17): 1809-1817	A vs. B (control)	Superiority	Two-sided	Superiority not confirmed	Both	Not reached due to early termination	Not stated at clinicaltrials.gov that study was terminated early.
106. JAMA; 2013; 310(17): 1818-1828	A vs. B vs. ∅	Superiority	Two-sided	Superiority not confirmed	Both	Reached as planned	Two cohorts: depressed and non-depressed patients. Two study arms for each cohort were compared to a control group. Two new methods were compared to a control

							group and were tested also for non-inferiority. ∅=control group
107. JAMA; 2013; 310(16): 1692-1700	A vs. Placebo	Superior-ity	One-sided α-risk of .025	Superior-ity not confirmed	Both	Not reached due to early termination	Clinicaltrials.gov: study has been stopped for fu-tility.
108. JAMA; 2013; 310(16): 1701-1710	A vs. AB New vs. Old	Superior-ity	Two-sided	Superior-ity not confirmed	Both	Reached as planned	
109. JAMA; 2013; 310(15): 1571-1580	A vs. AB New vs. Old	Superior-ity	Two-sided	Superior-ity not confirmed	Both	Reached as planned	Old=control
110. JAMA; 2013; 310(15): 1581-1590	A vs. B	Not stated {Superior-ity for A or B}	Two-sided	Not stated {Superior-ity not con-firmed}	Type-I-error	Sample size not described	
111. JAMA; 2013; 310(14): 1473-1481	A vs. ∅	Superior-ity	Two-sided	Superior-ity con-firmed	Both	Reached as planned	∅ = observation
112. JAMA; 2013; 310(13): 1353-1368	AB vs. Placebo A vs. Placebo	Not stated {Superior-ity}	Two-sided	Not stated {Superior-ity not con-firmed}	Type-I-error	Sample size not described	Two different study pop-ulations
113. JAMA; 2013; 310(12): 1240-1247	A vs. B New vs Old	Not stated {Superior-ity}	Two-sided	Not stated {Superior-ity con-firmed}	Both	Reached as planned	Old= control
114. JAMA; 2013; 310(11): 1135-1144	A vs. Placebo	Superior-ity	Two-sided	Superior-ity not confirmed	Both	Reached as planned	
115. JAMA; 2013; 310(11): 1145-1155	A vs. B New vs. Old	Superior-ity	Two-sided	Superior-ity not confirmed	Both	Reached as planned	
116. JAMA; 2013; 310(11): 1156-1167	A vs. ∅	Superior-ity	Two-sided	Superior-ity not confirmed	Both	Reached af-ter change of protocol	∅=no intervention
117. JAMA; 2013; 310(10): 1033-1041	AB vs. A + Placebo New vs. Old	Superior-ity	Two-sided	Superior-ity not confirmed	Both	Reached as planned	

118. JAMA; 2013; 310(10): 1042-1050	A vs. B vs. AB vs. ∅ New vs. Old	Superio-rity	Two-sided	Superior-ity not confirmed	Both	Reached as planned	B = practice-level incen-tives ∅=no incentives/inter-vention, control, old in-tervention
119. JAMA; 2013; 310(10): 1051-1059	AB vs. A New vs. Old	Superior-ity	Two-sided	Superior-ity not confirmed	Both	Reached as planned	
120. JAMA; 2013; 310(9): 918-929	A vs. B New vs Old	Superior-ity	Two-sided	Superior-ity con-firmed	Both	Reached as planned	Too many variables in the protocol (A consisted of two different treatment strategies, in which the physician decided the treatment regimen.) Old=usual care

Acknowledgements

I would like to thank my doctoral thesis supervisor Prof. Dr. Franz Porzsolt for broadening my horizon with this doctoral thesis and with the discussions we had. His constant advice, support, patience and availability helped me throughout the work.
Without him and the co-workers Karthik and Amit Ghosh, Oscar Kamga Wambo, Tania Gouvêa Thomaz, Cristiane Moraes, Valerio Balassone and Paola Rosati this project would not have been possible. I would like to thank them for their interest, time and patience during the project.

I would also like to thank Marie Ostermann, who helped me with professional advice and feedback and my aunt Juliane Brücker guiding me through the English grammar.

Another big help were my parents, my brother and my boyfriend. They supported and motivated me during all times and gave me strength to finish the work.

Curriculum vitae

Personal data:	Meret Phlippen
Born on:	02/11/1991
Born in:	Berlin
Parents:	Birgit Hammer-Phlippen, administrator
	Jörg Phlippen, director banking business
Siblings:	Lovis Phlippen (aged 26), studying mechanical engineering
Schooling	September 2003 until March 2010
	Bischöfliches Cusanus Gymnasium Koblenz
	Abitur: 1.8
	A-Levels: Biology, English, Latin
	September 2007 until March 2008
	Ratcliffe College Leicester
	A-Levels: Biology, History, Latin and Mathematics
Studies	October 2011 until November 2018
	Medical studies at the University of Ulm
	12th September 2013: First State Examination
	(Mark: 3.5)
	12th October 2017: Second State Examination
	(Mark 3.0)
	14th November 2018: Third State Examination
	(Mark: 2.0)
Work placements	Since February 2019:
	Assistant doctor for otorhinolaryngology in
	Vivantes Klinikum im Friedrichshain in Berlin