Threshold-Crossing for Single Arm Trials with External Control in Form of Aggregate Data

Threshold-Crossing für Einarmige Studien mit Externen Kontrollen in Form Aggregierter Daten

Masterarbeit “Master of Science”

im Studiengang “Angewandte Statistik”

an der Georg-August-Universität Göttingen

vorgelegt am 04. September 2022

durch Levin Wiebelt

Erster Gutachter: Prof. Dr. Tim Friede

Zweiter Gutachter: Prof. Dr. Tim Mathes

Table of Contents

Abbreviations

AGD: Aggregate Data 17
BAP: Bias-Adjustment Problem 5
ESS: Effective Sample Size 56
EU: European Union 14
EUnetHTA: European network for Health Technology Assessment 16
G-BA: Gemeinsamer Bundesausschuss 14
ICH: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use 6
Ide-Cel: Idecabtagen Vicleucel 5
IPD: Individual Participant Data 17
IQWiG: Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen 8
ITC: Indirect Treatment Comparison 11
MAIC: Matching-Adjusted Indirect Comparison 5
MM: Multiple Myeloma 18
ncp: non-centrality parameter 26
RCT: Randomized Controlled Trial 5
SAT: Single Arm Trial 5
TC: Threshold-Crossing 5
VAP: Variance-Adjustment Problem 5

1 Introduction

The use of *Single Arm Trials* (SATs) in pharmaceutical research is increasing (Patel *et al.*, 2021; Ribeiro *et al.*, 2022). However, applying SATs to generate evidence is a double-edged sword. On the one hand, it bears the potential of making research and drug development more efficient by producing faster results with less expenditure. On the other hand, it bears the risk of drawing more false conclusions from clinical trials. Authorities call for more rigor in the analysis of SATs (IQWiG, 2022b). Trial analysts and sponsors often face missing data problems. Multiple stakeholders in pharmaceutical research call for validation of new methodology to make new methods of evidence generation ready to apply in practice (Eichler *et al.*, 2020).

Following this call, this thesis investigates a framework for drug development programs called *"Threshold-Crossing"* (TC) (Eichler *et al.*, 2016), which centers around the concept of SATs. Its aim is to increase efficiency in pharmaceutical research, which may be achieved by arriving faster at decisions about the efficacy of the study drug.

The structure of the thesis is as follows. In Section 2 concepts of trial design are introduced. Differences between the gold standard design of the *Randomized Controlled Trial* (RCT) and the SAT design are explained. The problem of potential bias in the analysis of SATs becomes apparent. Section 3 discusses a recent rating of the evidence on a new drug called "Idecabtagen vicleucel" (Ide-Cel), where the relevant evidence consisted exclusively of SAT results. The bias problem outlined in Section 2 becomes salient in this example case. Section 4 introduces the statistical concepts of hypothesis testing, the t-distribution, as well as a derivative called the rescaled t-distribution, and a method for bias adjustment called *Matching-Adjusted Indirect Comparison* (MAIC). In Section 5, TC as described in the literature, is explained and its problems are pointed out. In this thesis these problems are categorized and coined as *Variance-Adjustment Problem* (VAP) and *Bias-Adjustment Problem* (BAP). A solution to the VAP is proposed. The BAP is a more fundamental problem in SAT designs. Nonetheless, addressing bias by MAIC is investigated. Section 6 outlines the conclusions of this thesis and discusses open points of debate.

2 Trial Design and Potential for Bias

2.1 Defining a Treatment Effect

In clinical trials the scientific interest is not only in association between treatment and endpoint, but in their causal relation. A *causal* treatment effect permits assertions like "improvement caused by the treatment". Associative assertions like "improvement followed by the treatment" are usually insufficient for assessing new drugs (IQWiG, 2020, p. 11). For estimating a causal effect, a comparison of patients' outcomes before and after treatment is applied is inappropriate. The patient's outcome may follow a systematic trend or may be due to the phenomenon of regression to the mean (Senn, 1997, chap. 3), which cannot be disentangled from the causal effect in a pre-post-comparison.

To understand the causal notion of a treatment effect it is useful to think of the counterfactual outcome of a patient, which would result had no treatment been given (Rubin, 1974). By definition, the counterfactual outcome is unobservable. Hence the counterfactual is required to be estimated. This is usually done by not treating all patients with the study drug, but assigning some patients to the control group. The control group can be used to estimate the counterfactual for the treatment group. Performing the comparison on a group level is in line with the principles of evidence-based medicine, stating that statements on individual patients are impossible, while statements on groups of patients are possible (IQWiG, 2022a, p. 7).

In practice, different control conditions or counterfactuals may be of interest. The simplest one may be refusing treatment. In some settings the control condition of interest may be giving a *placebo*, which is a pseudo-treatment without true biological effect. Placebos are used to eliminate the influence of knowledge about the treatment status by patient or investigator (Piantadosi, 2005, sec. Appendix B.3). Further control conditions are an alternative drug under investigation or the standard of care.

The scientific question of interest, including the control condition, can be operationalized by defining an estimand. The *International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use* (ICH) defines an estimand as a precise description of the treatment effect, constituted by five attributes (ICH E9 R1, 2019): Population, Endpoint, Summary Measure, Treatment and

Intercurrent Events. Of special interest in this thesis will be two of the attributes: the specification of the patient *population* and the *summary measure* for comparing patient observations of the respective trial arm. The definitions of the *endpoint* and of the *treatment* will be assumed as given in the following. Strategies for handling *intercurrent events* will neither be investigated in this thesis. If correctly specified, the estimand corresponds to the causal effect, which is of scientific interest. The conduct of the trial will focus in large parts on estimating this quantity.

2.2 Bias

Bias is a property of a point estimator $\hat{\delta}$ for an estimand δ. It denotes the difference between the expected value of the estimator and the estimated quantity or estimand: $bias(\hat{\delta}) = E[\hat{\delta}] - \delta$ (Casella and Berger, 2002, chap. 7). Bias is the systematic deviation of the estimation result from the unknown truth.

To illustrate the concept of bias, consider the causal quantity of interest in a trial is the effect of receiving treatment compared to receiving placebo. Imagine a trial, that failed to assign placebo as control condition. Patients know about their treatment status, as well as investigators, which gives rise to a possible placebo effect. Estimating a treatment effect by comparing treatment and control group yields an effect estimate, that is an intermingled quantity, consisting of the causal treatment effect and the placebo effect. However, the relative magnitudes are unknown and not estimable in this setting. Consider a moderately sized positive effect estimate. The causal effect may constitute a large share of the estimate, while the placebo effect constitutes a small share of the estimate. However, it may also be vice versa. The causal effect is relatively small, while the placebo effect is relatively large. Given a positive effect estimate, the causal effect may even be zero or negative, masked by a large placebo effect.

Based on the effect estimate, multiple magnitudes of the true causal effect are plausible, depending on the unknown amount of bias. These different magnitudes would yield qualitatively different trial results. Hence the effect estimate comes with uncertainty in the presence of potential bias. Bias must be addressed already prior to statistical analysis by a good trial design (Piantadosi, 2005, chap. 7). Interest in clinical research is often in small treatment effects (Piantadosi, 2005, chap. 2). This makes the task of disentangling causal effect from biasing factors especially important, since

already small bias has a substantial influence in the effect estimate and can drastically change the conclusion of the trial.

A systematic way to assess the risk of bias of a study requires a classification of different bias sources. The ROBINS-I tool (Sterne *et al.*, 2016) is a framework targeting risk of bias assessment specifically in intervention studies. The tool categorizes potential bias into different bias sources, that may arise at different stages of the trial, namely pre-intervention, at-intervention and post-intervention. Of special importance in this thesis are the bias sources arising at the pre-intervention stage: *selection bias* and *confounding bias*, which are explained in the subsections below. The ROBINS-I tool is used for orientation in assessing the bias potential of a trial by the "*Institut für Qualität und Wirtschaftlichkeit im Gesundheitswesen*" (IQWiG) (IQWiG, 2022a, p. 172). The IQWiG is a scientific institute, that regularly performs evidence ratings of clinical trials in Germany to support decisions within the national health care system.

2.2.1 Confounding

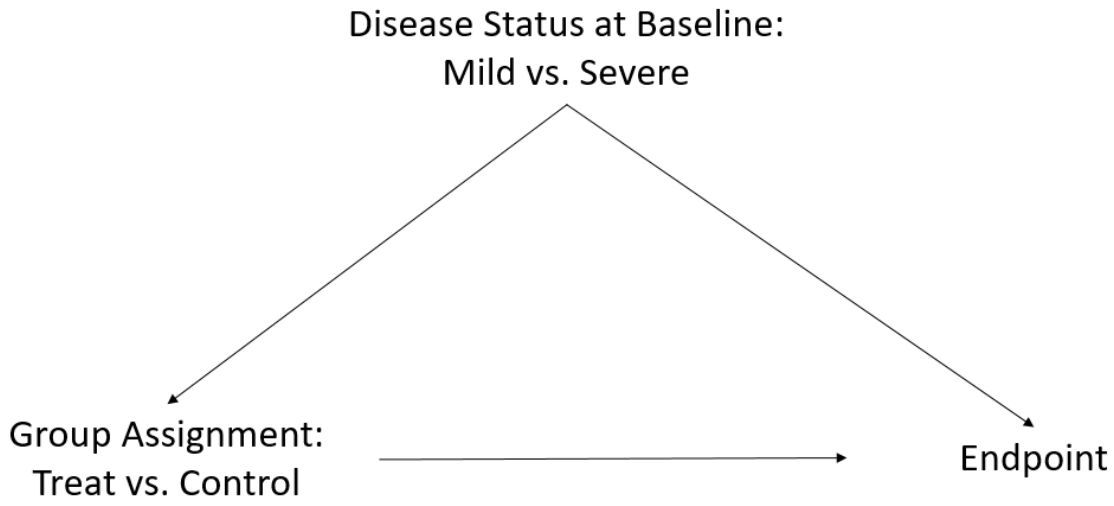


Figure 1: Confounding
An arrow in this graph represents a causal effect of one variable on another. The direction of causality is indicated by the tip of the arrow. Group Assignment has a causal influence on the Endpoint. Simultaneously, Disease Status has an influence on both Group Assignment and the Endpoint. Differences in the Endpoint between groups cannot be interpreted to be caused by Group Assignment, since the change may be due to the difference between the groups with respect to the confounder "Disease Status". The causal effect of Group Assignment on the Endpoint cannot be estimated unbiased.

Confounding occurs if an interference factor of the design has influence on the assignment of patients into the trial arms, as well as on the measured outcome variable (Hernán, 2014). Consider the case where patients may have a mild or severe disease status at baseline. Let the endpoint be some measurement of the disease status after

treatment has been applied. Likely, the interference factor “Disease status at baseline” has an influence on the endpoint. Let “Disease status at baseline” be encoded as a binary variable. Assume the distribution of disease status differs between the groups. For example, the control group contains a larger share of severely diseased patients, whereas in the treatment group the share of mildly diseased patients is higher. In this case the baseline variable not just associated with the endpoint, but additionally with treatment assignment. Therefore, it confounds the estimation and a potentially biased effect estimate results. A causal diagram similar to (Tennant *et al.*, 2021) and following the conventions of (Pearl, 1995), the confounding scenario looks as displayed in Figure 1.

2.2.2 Selection Bias

If there are systematic differences between trial arms with respect to the definition of the inclusion and exclusion criteria of patients, the resulting bias is referred to as *selection bias* by the ROBINS-I-tool. An example is comparing a group of prevalent users of a treatment to a control group (Sterne *et al.*, 2016). The correct design is comparing new users of a treatment to the control group. The reason is that patients who initially do not respond well to the treatment tend to select out of the group of prevalent users. This makes prevalent users systematically better off than new users, which overestimates the treatment effect of interest.

2.3 Dramatic Effects

If the causal treatment effect is large, bias is of less concern. The reason is that small or moderate bias does not qualitatively change the trial result in the presence of a large causal effect. Such a setting is referred to as “*dramatic effect*” (IQWiG, 2022a, p. 59 f.). If an exceptionally large treatment effect is estimated, it is deemed implausible to be caused exclusively by biasing factors. In this case it is deemed plausible, that at least some part of the possibly biased large effect estimate is due to a causal treatment effect.

The IQWiG refrains from following a strict definition of dramatic effects. However, for orientation they consider an estimated relative risk of 10 accompanied by a significance level of 1% as an indication for a dramatic treatment effect (IQWiG, 2022a, p. 59 f.). In these cases, the IQWiG may consider even a pre-post comparison of patients as

sufficient evidence for its decisions. At the same time, it points out that dramatic treatment effects sizes are very rare in modern medicine.

2.4 Randomized Controlled Trial (RCT)

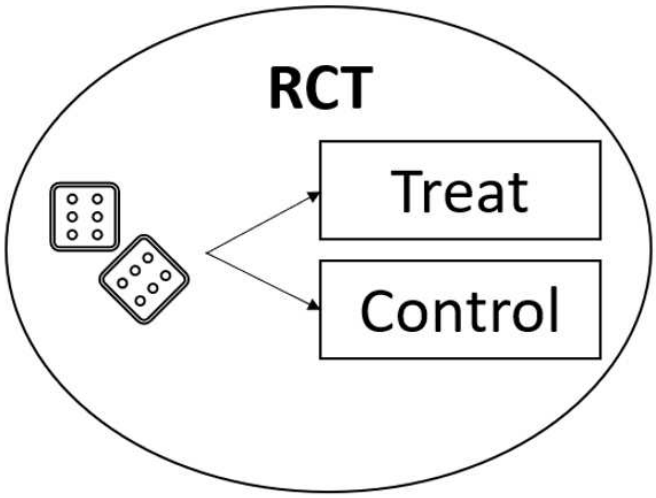


Figure 2: RCT Design
In the RCT design recruited patients get assigned into treatment group and control group by a random allocation process.

A trial design that minimizes the potential for bias is the RCT. It has two central features:

- Parallel Control Units, and a
- Randomized Group Assignment.

The partition of a single patient cohort intro treatment and control arm ensures that the same population criteria apply to both arms, which rules out selection bias. The randomized assignment of recruited patients into treatment arm or control arm of the trial eliminates systematic differences between trial arms with respect to baseline characteristics of the patients. This holds for observable differences in terms of measured baseline characteristics, but also for unobserved characteristics (Piantadosi, 2005, chap. 3). If no systematic differences between groups exist, there is no factor influencing the group assignment and therefore no systematic confounding. Note that due to the random group assignment, there may be baseline differences between arms which are due to random deviations between patients.

The minimization of bias sources qualifies the RCT design for investigating medical treatment with a small effect size, that would be masked if bias was present.

2.5 Single Arm Trial (SAT)

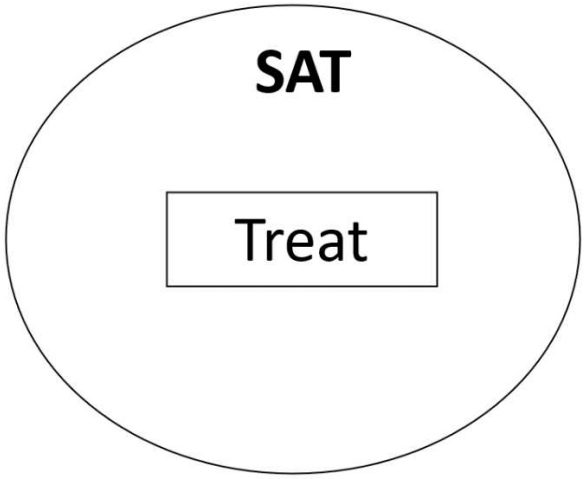


Figure 3: SAT Design
In the SAT design all patients recruited are treated by the study drug. This implies that only a treatment group is available, while a control group is missing,

A Single Arm Trial is conceptually simpler than an RCT. All patients recruited receive the treatment under investigation. This implies that the trial consists of a single treatment arm, a control arm is missing. A pre-post comparison would be possible based on the treatment arm only. This is prone to bias. It cannot disentangle the causal treatment effect from systematic time trends of the disease (Senn, 1997, chap. 3).

Another possibility for making a comparison based on a SAT is to use an external control cohort. For instance, it may be a patient cohort of another trial investigating a similar patient population or a patient registry. This corresponds to the idea of an unanchored Indirect Treatment Comparison (ITC) (Bucher *et al.*, 1997). A SAT performed like this is missing the two central features of the RCT, namely:

- parallel control units, and
- randomized group assignment.

This gives rise to the potential for bias in the SAT design. To assess it, a rigorous assessment of the patient groups must be performed. Following the ROBINS-I tool confounding and selection bias must be assessed at the pre-intervention stage. Selection bias is assessed by ensuring that the same criteria for including and excluding patients apply for the study-internal treatment arm and for the study-external control arm. Usually this is done by subsetting the external control group to eligible patients fulfilling the treatment group population criteria. The risk of confounding can be

addressed by systematically determining all relevant confounders for the comparison. The possibility of adjusting for them depends on the data availability of confounding information in the trial arms. Measured confounding can be adjusted for, the risk of unmeasured confounding eventually remains unaddressed.

Depending on the success of mitigating bias one of the two scenarios illustrated in Figure 4 can result. On the left-hand side a successful application of a SAT is illustrated. Selection bias is addressed by using the same population criteria for treatment and control arm. This makes the trial arms structurally comparable. Still, confounding bias must be addressed in the statistical analysis. This is required since patients are not randomly assigned to treatment and control group, but follow an unknown allocation process as illustrated by the question mark. Adjusting for confounding bias requires a model for the unknown allocation process in the statistical analysis. In many cases this is done by using so called *Propensity Scores* (Austin, 2011). In this thesis, however, another method called MAIC (Signorovitch *et al.*, 2010) is used for confounder adjustment. MAIC is designed to be applicable even if data availability of the external control group is limited.

If mitigating bias in a SAT fails, the right-hand side illustration in Figure 4 is more adequate. In some cases it may be impossible to apply the same population criteria to the two trial arms, which makes them structurally incomparable. In some scenarios structural comparability may be given, but missing data on confounding factors prohibits the adjustment for confounding bias, which also results in an invalid comparison of the trial arms.

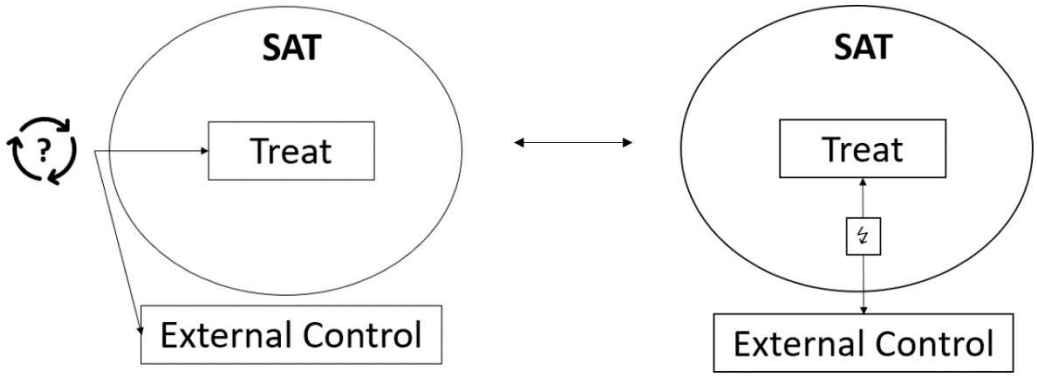


Figure 4: Valid (left) and Invalid (right) SAT
On the left-hand side a valid SAT is illustrated. Trial arms are structurally comparable. In contrast, to the RCT patients are not assigned to the groups via a random allocation mechanism. The allocation mechanism in a SAT is unknown. This requires the modelling of the unknown process prior to treatment effect estimation. On the right-hand side an invalid SAT is illustrated. The trial arms are incomparable, which may be due to selection or confounding bias. An unbiased treatment effect cannot be estimated.

Even if selection and confounding bias are addressed properly, the SAT still lacks some characteristics of a RCT. Treatment allocation can usually not be concealed. This makes patients and physicians prone to confirmation bias, which results if the endpoint measurement is influences their expectations. For some endpoints it may be possible to blind the investigator (Ford and Norrie, 2016), for instance if the endpoint is radiological imaging. An option to mitigate the influence of allocation knowledge is choosing an objective endpoint, such as “overall survival” or “emergency hospital admissions” (Ford and Norrie, 2016).

Using patient cohorts of different data sources demands checking for coherent definitions of variables and coding schemes between the sources. Hernán and Robins set the example of using a database of health care claims as external data source (Hernán and Robins, 2016). Let patients with breast cancer be the relevant population. A breast cancer diagnosis in the external data base may represent a true diagnosis for some patients, while other breast cancer diagnosis may represent a suspicion of the physician, which is recorded in order to perform more diagnostic tests (Hernán and Robins, 2016). Applying this population criterion to the external database in this example requires a preceding validation of the variable definition.

If there is a time lack in observing treatment and control arm, a changing standard of care may bias the trial result. This is of special concern in indications with rapid development of treatments. For instance, in the 1990s and 2000s the standard of care in treating AIDS has been improving rapidly, such that historical controls would have been inadequate as control group (Piantadosi, 2005, chap. 2). The reason is that historical controls are systematically worse-off due to the lower standard of care. A comparison with the treatment group would overestimate the causal treatment effect.

In order to minimize these further bias sources, the idea of emulating a target trial is very useful (Hernán and Robins, 2016). The idea has been developed for application in observational database studies. It can be of equal use in single arm intervention studies, that use observational data as control group. (Hernán and Robins, 2016) call for framing the scientific question as one that would be answered by a RCT, while subsequently emulating the RCT.

Designing a trial as single armed can yield advantages in certain research settings. If the treatment under investigation is highly promising in easing or even curing a disease, the assignment of patients to the control group may be ethically questionable. Additionally, patients or physicians may be unwilling to consent to randomized treatment assignment. This consideration is even more serious if the disease to be treated is known to have a severe progression. Not assigning control condition does not raise the ethical problems of refusing treatment. Additionally, a single treatment group can be populated much faster than populating two trial arms with patients.

2.6 Benefit Assessment

In the European Union (EU), decisions on marketing authorization of a drug are made centralized by the European Medicines Agency (EMA). The process of benefit assessment starts after the approval process. In Section 3 an example of a benefit assessment process is discussed. Benefit assessment is not centralized, but organized nationally in the European Union. Benefit assessment serves as the basis of price setting of drugs, as well as for deciding whether the national health system funds the drug or not. For a detailed description of this process see (Leverkus and Chuang-Stein, 2016). The authority in Germany for deciding on the benefit of a drug is the "Gemeinsamer Bundesausschuss" (G-BA). The IQWiG is a research institute, which has been awarded a general mandate by the G-BA for several tasks including benefit assessment.

Benefit assessment concerns the comparison of a new drug to the best available alternative. Hence, not only efficacy, but also efficiency of a treatment over another treatment is of interest. The evidence is usually generated by the drug manufacturer conducting clinical trials and describing evidence in a submission to the G-BA or IQWiG. On the basis of this submission the authority performs an evidence rating and decides on the benefit of the study drug. Guidelines for evidence rating are the code of procedure of the GBA (G-BA, 2008) and the general methods of the IQWiG (IQWiG, 2022a).

2.7 Rating Evidence

Benefit assessment requires evidence rating. Rating evidence is a complex task. There is no scientific consensus on which rating scheme to use. Note that the rating schemes of the GRADE working group receive lots of support and claim theirs being the new consensus (Guyatt *et al.*, 2008). The IQWiG, in contrast, uses a conventional hierarchy of study designs for orientation. They do not describe their scheme explicitly but refer to the scheme used by the G-BA (G-BA, 2008), which matches in large parts with the conventional pyramid scheme for evidence rating displayed in Figure 5.

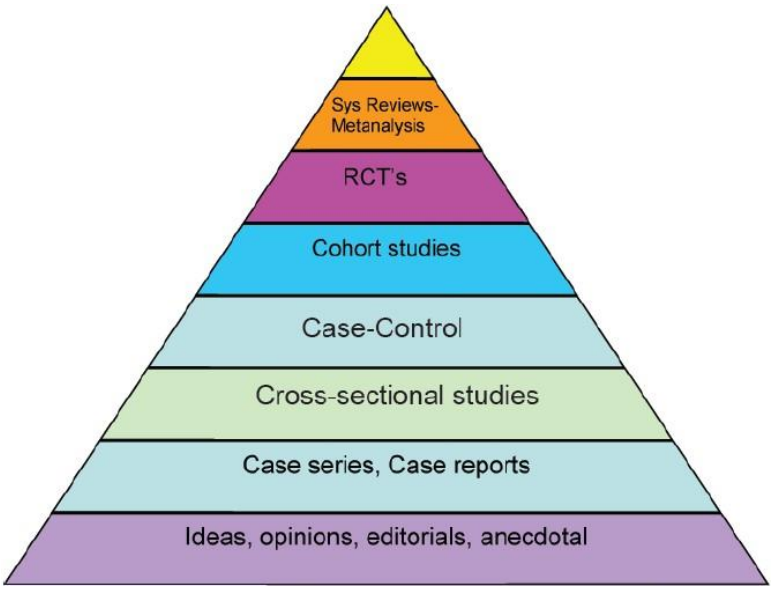


Figure 5: Evidence Hierarchy (Phillips, 2014)
Highest evidence quality in this scheme draws on systematic reviews and meta-analysis, which are summaries of multiple individually conducted trials. RCTs possess highest evidence rating, followed by non-interventional study designs, such as cohort, case-control, and cross-sectional studies. Case series and expert opinions are located at the bottom of the pyramid, illustrating the lower evidence ratings.

The highest evidence quality is deemed for review studies, that summarize individual trials. These are meta-analysis, which are not discussed in this thesis. Of individual trial designs the RCT is deemed to have best evidence quality. This is due to the low potential for bias as explained above. In contrast to RCT and SAT as described above, cohort, case-control and cross-sectional studies are no intervention studies, which makes these designs unusable for estimating the effect of a new medical treatment. Case series are located close to the bottom of the pyramid scheme, indicating low evidence quality.

The conventional pyramid scheme fails in rating evidence quality of non-randomized intervention trials, such as a SAT. The evidence quality of non-randomized trials is largely dependent on its potential for bias. As mentioned above, the IQWiG points out

that in case of dramatic effects also case series may be used as a control in an intervention trial (IQWiG, 2022a, p. 59 f.). This fact should make clear that this hierarchy is not strictly applied but the evidence rating is very much dependent on the concrete research setting.

A very useful way to think about evidence quality is suggested by the *European network for Health Technology Assessment* (EUnetHTA). Evidence quality can be thought of in three dimensions: internal validity, statistical precision and external validity (EUnetHTA, 2021).
Up to this point this thesis discussed points considering the *internal validity* of a trial, that is the potential for bias and comparability of trial arms. Low internal validity leads to systematic errors in the results drawn from a trial. However, even in the absence of systematic errors or bias, random errors may still occur in the analysis of a trial (Piantadosi, 2005, chap. 7). Addressing the rate or probability of random errors can be summarized by the concept of S*tatistical precision*. The two-sample t-test, a statistical method to quantify random error rates, is introduced in Section 4. *External validity* captures the notion of transferability of trial results to clinical practice. This concept is not of central concern in this thesis, but common arguments on external validity are discussed in Section 6.

2.8 Categorizing External Data

Multiple sources for external control cohorts are available to compare the SAT treatment arm against. One option is to use the control group of a past RCT cohort. This may come with limited information on possible confounding variables since confounding bias is of less concerns in RCTs than in non-randomized designs. RCTs are interventional trials, in which a treatment is experimentally applied. If the study drug is to be compared against the best available alternative, the treatment arm of the approval trial of the best alternative may yield a good comparator. Cohorts from current or past cohort studies or case-control studies, in contrast, are non-interventional. Data taken from studies can be classified as “Research Data”, following the scheme of Franklin and Schneeweiss (2017).

On the other side of this scheme is transactional data, which arise automatically if patients make use of the national health system (Franklin and Schneeweiss, 2017). Data of this category is often referred to as “Real World Data”. Use of “Real World

Evidence" is deemed to increase external validity, which is discussed in Section 6 of this thesis. Examples of transactional data are patient cohorts constructed from health insurance claims or from patient-generated data via health apps. Often times, only surrogate endpoints are available, requiring validation studies in advance of using transactional data (Chen *et al.*, 2021). It is deemed of lower quality than research data.

Patient registries vary in their quality. Some can be categorized as research data, some rather as transactional data. Some registries even bear the potential of being a complete survey of the patient population (Chen *et al.*, 2021).

An important distinction regards the availability of the data to the trial conductor. The external data may be fully available as individual-participant-data (IPD) or more restricted in the form of aggregate data (AGD), which could be for example summary statistics from the publication of a past trial. All methods discussed in the following work under the restricted availability of external control data as AGD. Note that data for the treatment group is actively collected within the SAT and therefore is available as IPD.

3 SATs in Benefit Assessment: Example Case “Ide-Cel”

In this section, the benefit assessment of a drug called Idecabtagen Vicleucel (Ide-Cel) for the German health system is presented. It was performed by the G-BA and published in April 2022. A courtesy translation exists (G-BA, 2022a), however only the German version (G-BA, 2022b) is legally binding. The Ide-Cel case exemplifies many considerations explained in Section 2 of this thesis, especially the difficulties of generating evidence by the SAT design.

3.1 Patient Population

Ide-Cel is a treatment for multiple myeloma (MM), a cancer disease of the plasma cells, a type of white blood cells. The patient population targeted by this treatment consists of “adults with relapsed and refractory multiple myeloma who have received at least three prior therapies, including an immunomodulatory agent, a proteasome inhibitor and an anti-CD38 antibody and have demonstrated disease progression on the last therapy.“ (G-BA, 2022a) Due to this narrow selection, a small patient population with relapsed or refractory multiple myeloma results.

3.2 Result of Benefit Assessment

The GBA’s final judgement is granting an non-quantifiable benefit of Ide-Cel (G-BA, 2022b). However, the decision was not based on the submission by the drug manufacturer, but by the so called “orphan drug” status (EMA, 2018) of Ide-Cel. For orphan drugs, benefit is granted based on the earlier approval of the drug for marketing authorization.

Nonetheless the manufacturer’s submission was assessed by the G-BA to eventually quantify the benefit of Ide-Cel. The manufacturer’s submission consisted of evidence based on SATs and was considered unsuitable for quantifying the benefit.

The main argument for rejecting the evidence of the submission is the incomparability of trial arms. The comparisons presented in the submission are indirect, which means that trial arms consist of cohorts from different studies. For indirect comparisons to be valid, the analysis must ensure the trial arms to be both

- structurally sufficiently similar, and
- adjusted for all relevant confounders (G-BA, 2022b).

Note that this argument can be illustrated with categories of ROBINS-I at the pre-intervention stage described in Section 2.2: selection bias and confounding bias.

Selection bias arises due to differences between trial arms with respect to inclusion and exclusion criteria of patients, while confounding occurs if differences exist with respect to baseline characteristics.

3.3 Orphan Drug Status

The orphan status is granted by an EMA-committee for medicinal products fulfilling certain criteria. An overview can be found on the EMA-website (EMA, 2018), while the legally binding definition is found in the corresponding EU regulation (European Parliament, 1999). The criteria may be summarized as follows:

- The targeted disease is life-threatening or chronically debilitating
- Either the prevalence of the disease is 0.05% or less in the EU or the probability of marketing revenue justifying the investment in drug development is low
- No satisfactory alternative medicinal product exists.

Note that the orphan drug status can still be granted if the third condition is not fulfilled, that is alternative products exist. However, in this case the product under investigation must prove a benefit over the alternative. In the case of Ide-Cel the benefit argument went the conventional way: orphan drug status was granted and based on this the benefit of the drug.

3.4 Safety Assessment

Assessment of safety is rejected by the GBA. Non-adjusted comparisons of incidence rates of adverse events were presented by the manufacturer. Effect estimators were not calculated. Without adjustment sufficient structural equality of the patient populations cannot be assumed.

3.5 Efficacy Assessment

The evidence on efficacy considered by the G-BA consists in three ITCs based on four patient cohorts, which are the treatment cohort called “KarMMa” and three external control cohorts:

- NDS-MM-003 (in the following NDS),
- PREAMBLE, and
- MM-007 (in the following OPTIMISMM)

The KarMMa cohort is extracted from the SAT used for approving Ide-Cel for marketing authorization by the EMA. The NDS-cohort is collected retrospectively from clinical centers and research databases. The PREAMBLE cohort is collected prospectively in

different study sites including university hospitals and doctor's practices. The OPTIMISMM-cohort is taken from an RCT. Three unanchored ITCs are performed, comparing KarMMa to each one of the external control cohorts.

Further comparisons were included in the manufacturer's submission, which were judged as irrelevant for benefit assessment by the GBA. These consisted in an ITC of the patient cohort from the "CRB-401"-trial, which was a supplementary study in the approval process of IdeCel. The GBA criticized that the estimate for the endpoint overall survival, which is central for efficacy assessment, is invalid.

Each ITC performed consists of two steps. In a first step, "eligible" patients of the external control cohort are subsetted from the cohort. Eligible patients match the population criteria of the KarMMa trial as much as possible. This step aims at reducing differences with respect to population criteria between trial arms, that is reducing selection bias. The KarMMa-population is roughly defined as patients with MM that received at least three prior therapies and are refractory to the last. For more detailed inclusion and exclusion criteria see (G-BA, 2022b). The second step consists in statistical analysis, that adjusts for available information on confounders. Data on the external control cohorts were available as IPD. This is why the trial conductors chose Propensity Score methods for confounder adjustment in the analysis of the trial, which is considered adequate by the G-BA (G-BA, 2022b). Propensity Score methods are not considered in this thesis. Instead the MAIC method for confounder adjustment is described and investigated in Section 5. MAIC is applicable in case of AGD-availability of the external control cohort, while Propensity Score methods are a common choice in case of IPD availability.

3.5.1 Determining Relevant Confounders

Prior to data analysis, relevant confounders were required to be determined by the manufacturer. This was performed by a two-track strategy consisting of systematic literature research and expert interviews. The G-BA questions the completeness of the list of relevant confounders presented by the manufacturer. This critique is based on differences in results between literature research and expert interviews and on shortcomings of the literature research. The G-BA doubts that the literature research has been too narrow. The research consisted in screening 63 publications, including

meta-analyses, ITCs and other studies. Confounders used in these studies were collected and reported. The G-BA finds that this procedure is limited to finding only confounders with data availability. Lastly, the G-BA finds inconsistencies in the resulting list of relevant confounders as presented by the manufacturer and the sources used to determining them (G-BA, 2022b).

The critique of limiting confounder research to ones with data availability is an interesting one, since confounders without data availability could not possibly be adjusted for. Reporting relevant confounders without data availability would create problems further down the pipeline of evidence generation. Obviously, the manufacturer is not able to adjust the analysis for confounders without data availability. An analysis unadjusted for relevant confounders would also be unacceptable for the G-BA.

3.5.2 ITC – KarMMa vs. NDS

The NDS-cohort includes past trial data, as well as data from scientific databases, which is collected retrospectively. It may be the case that individual patients may have entered both cohorts NDA and KarMMa simultaneously. The reason is overlap in the enrolment of subjects, which applies to 10 percent of KarMMa patients and 23 percent of NDS patients. However, due to these low shares the bias arising due to overlap is considered insignificant (G-BA, 2022a). Considering confounding bias, the G-BA criticizes that some confounders identified as relevant by manufacturer did not enter the analysis. confounders with a share of more than 30 percent missing data were not included.

3.5.3 ITC – KarMMa vs. PREAMBLE

The PREAMBLE cohort is taken from a prospective cohort study aiming to investigate everyday clinical health care of MM patients. An exclusion criterion is the enrollment of the patient in a clinical trial. Hence recruitment-overlap with KarMMa is no concern. Data on country and setting in which patient recruitment took place (clinical center, family practice, etc.) is missing. Potential for selection bias arises in the process of matching KarMMa-population criteria due to missing data. Potential for confounding bias is present as well. Some confounders do not enter analysis due to missing data. Additionally, the distribution of “cytogenetig risk”, a confounder that entered analysis, is not reported for the trial arms after confounder adjustment.

3.5.4 ITC – KarMMa vs. OPTIMISMM

OPTIMISMM is a RCT cohort. Applying the KarMMa population criteria would result in few eligible patients for comparison. The manufacturer solved this problem by applying the KarMMa criteria in a less strict way. Patients who fulfill the criteria not yet at baseline, but eventually during the follow-up period are also included in the comparison. The problem was mainly caused by the KarMMa-criterion of having received at least three prior treatments. The G-BA states that this approach is adequate in principle, while raising the risk of bias (G-BA, 2022b). Potential for confounding bias arises in this ITC as well, due to the missing adjustment for some relevant confounders.

3.6 Summary

The main argument for excluding the evidence on efficacy of Ide-Cel based on the unanchored ITCs presented above is the potential for bias. As mentioned above for a valid unanchored ITC trial arms need to be both

- structurally sufficiently similar, and
- adjusted for all relevant confounders (G-BA, 2022b).

The former point refers to selection bias, while the ladder refers to confounding bias. For either ITC it is possible that patients in the highly selected KarMMa-cohort are systematically better-off than patients of the less selected external control cohorts, which would lead to an overestimation of the treatment effect. Matching the KarMMa population criteria is limited due to missing information and insufficient reporting by the manufacturer. It cannot be excluded that KarMMa patients are, due to narrow selection criteria, systematically better-off than patients of the control cohorts. Missing data prohibits the adjustment for identified relevant (measured) confounders. Additionally, the G-BA questions if the list of relevant confounders determined by the drug manufacturer is complete. This gives rise to the potential for unmeasured confounding. These bias concerns would be less grave in the case of dramatic effects. However, estimated treatment effects are moderate, and bias concerns receive a high weight in the judgement. Therefore, a quantification of the benefit is not possible based on the evidence presented in the submission (G-BA, 2022b).

4 Statistical Foundations

Error in trial results can be categorized as *systematic* and *random error* (Piantadosi, 2005, chap. 7). Section 2 introduced concepts of bias and design to assess systematic error in a trial. This section, in contrast, introduces the statistical concepts to address random error in a trial. Following the categories suggested by the EUnetHTA, addressing systematic error corresponds to judging *internal validity* of a trial, while addressing random error corresponds to judging *statistical precision* of a trial (EUnetHTA, 2021).

4.1 Hypothesis Tests

A hypothesis test is a statistical framework for making binary. The decision may for instance be whether a study drug is efficacious or not. Two competing statistical hypotheses are stated, which are referred to as null (H_0) and alternative hypothesis (H_1) (Piantadosi, 2005, chap. 7). In the following, one-sided hypotheses are formulated:

$$H_0: \quad \mu_{treat} - \mu_{hist} \leq 0; \quad H_1: \quad \mu_{treat} - \mu_{hist} > 0$$

Initially, the null hypothesis is assumed to be true. The plausibility of this assumption is assessed by a test statistic, which is a summary of the available data (Piantadosi, 2005, chap. 7). To quantify the probability of observing a certain test statistic value under H_0, the distribution of the test statistic under the null must be known or approximable. A decision boundary is set in advance based on this distribution. If a test statistic beyond the decision boundary is observed, the null hypothesis is judged implausible and is rejected.

A hypothesis test allows the quantification of random decision error. Type-I- and type-II-error are distinguished (Piantadosi, 2005, chap. 7). A type-I-error or false positive occurs if the null is true, but falsely rejected. This corresponds to judging an inefficacious drug as efficacious. A type-II-error or false negative occurs if the alternative is true, but the null is falsely retained This corresponds to judging an efficacious drug as inefficacious.

The probability of type-I-error is required to be controlled in a trial, for instance the G-BA requires a significance level of $\alpha = 2.5\%$ for one-sided hypothesis tests (G-BA, 2008). The significance level corresponds to the upper bound for the permissible probability of type-I-error (Searle and Gruber, 2016, chap. 3). If the distribution of the

test statistic under the null hypothesis is known, the type-I-error probability can be controlled by setting the respective $(1-\alpha)$-quantile as a decision boundary. This is the case for the two-sample t-test, which is described in the following section.

In contrast, addressing the probability of type-II-error is rather in the trial sponsor's interest (ICH E9, 1998, p. 18). It is more convenient to refer to the *power* of a trial, which is the complement of type-II-error probability, in the sense that both add up to one. In the following parts of this thesis the power target is set to 80%, following the specification in Eichler *et al.* (2016). The first step in powering a trial is to formulate the alternative hypothesis explicitly. This involves consultation of experts on the question of which effect size is of clinical relevance. The minimum clinically relevant effect size δ^* is used:

$$H_1: \quad \mu_{treat} - \mu_{hist} = \delta^*$$

The distribution of the test statistic under the alternative is different than under the null. It is dependent on the sample size of the trial. Powering a hypothesis test corresponds to collecting sufficient datapoints. In the tests discussed in this thesis closed form sample size formulae are available. These are useful approximations of the observations needed to achieve the power target of 80%.

4.2 Two-Sample T-Test

A t-test arises if means are compared between two groups, where endpoints are normally distributed. Consider a model where observations scatter with equal measurement variance σ^2 around a group-specific mean μ_i (Büning, 2002).

$$y_{i,j} = \mu_i + \varepsilon_{i,j} \sim N(\mu_i, \sigma^2); \qquad i = treat, hist; \; j = 1, \dots, n_i$$

For constructing the test statistic sample means are required. Since the sample mean is a linear combination of normally distributed variables, it is itself normally distributed (Büning 2002):

$$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{i,j} \sim N\left(\mu_i, \sigma^2 * \frac{1}{n_i}\right); \qquad i = treat, hist; \; j = 1, \dots, n_i$$

Given the independence of the samples, the distribution of the sample mean difference is as follows:

$$\bar{y}_{treat} - \bar{y}_{hist} \sim N(\mu_{treat} - \mu_{hist}, \sigma^2 * (\frac{1}{n_{treat}} + \frac{1}{n_{hist}}))$$

A standardization by the correct variance would result in a standard normally distributed quantity under the null:

$$Z = \frac{\bar{y}_{treat} - \bar{y}_{hist}}{\sigma\sqrt{\frac{1}{n_{treat}} + \frac{1}{n_{hist}}}} \overset{H_0}{\sim} N(0,1)$$

The random variable Z contains the unknown population parameter $\sigma = \sqrt{\sigma^2}$. An estimate for the measurement variance σ^2 is needed. In case of equal variance across groups a pooled estimate can be used (Büning 2002):

$$S^2_{pooled} = \frac{1}{N-2}\sum_i \sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_i)^2; \qquad N = n_{treat} + n_{hist}$$

Note that a pooled estimate can be calculated from group-wise variance estimates by the following formula (Büning 2002):

$$S^2_{pooled} = \frac{(n_{treat} - 1) * S^2_{treat} + (n_{hist} - 1) * S^2_{hist}}{N-2},$$

$$where \quad S^2_i = \frac{1}{n_i - 1}\sum_{j=1}^{n_i} (y_{i,j} - \bar{y}_i)^2$$

The ratio of the sample variance and the population variance multiplied by the degrees of freedom used in estimation is chi-square-distributed, with the respective degrees of freedom (Rencher and Schaalje, 2008, chap. 7):

$$V = (N-2)\frac{S^2_{pooled}}{\sigma^2} \sim \chi^2(df = N-2)$$

A central t-distribution arises for a ratio term that includes, both Z and V, given their independence (Johnson, Kotz and Balakrishnan, 1995, chap. 28). Precisely, the term corresponds to the standardized means difference Z, divided by the root of the chi-square distributed variable V, divided by its degrees of freedom:

$$T = \frac{Z}{\sqrt{\frac{V}{df}}} \overset{H_0}{\sim} t(df = N-2)$$

If the terms for Z, V and df are inserted, T reduces to the sample mean difference $\bar{y}_{treat} - \bar{y}_{hist}$ divided by its standard error $\widehat{se}(\bar{y}_{treat} - \bar{y}_{hist})$. T corresponds to the test statistic of the two-sample t-test.

$$T = \frac{Z}{\sqrt{\frac{V}{df}}} = \frac{\bar{y}_{treat} - \bar{y}_{hist}}{\sigma\sqrt{\frac{1}{n_{treat}} + \frac{1}{n_{hist}}}} * \sqrt{\frac{(N-2)*\sigma^2}{(N-2)*S^2_{pooled}}} = \frac{\bar{y}_{treat} - \bar{y}_{hist}}{\widehat{se}(\bar{y}_{treat} - \bar{y}_{hist})},$$

$$where \quad \widehat{se}(\bar{y}_{treat} - \bar{y}_{hist}) = S_{pooled}\sqrt{\frac{1}{n_{treat}} + \frac{1}{n_{hist}}}$$

Using the $(1-\alpha)$-quantile of the t-distribution as decision boundary results in a constant type-I-error probability under the model assumptions:

$$T^{crit} = t^{-1}(1-\alpha, df = N-2); \qquad P\left(T > T^{crit} | H_0\right) = \alpha$$

Power and the probability of type-II-error are conditional probabilities given the alternative:

$$P\left(T > T^{crit} | H_1\right) = 1 - t(T^{crit},\, ncp = \vartheta)$$

In case the alternative distribution is true, the test statistic T calculated as above follows a non-central t-distribution. It is additionally parametrized by a non-centrality parameter (ncp) ϑ. The expectation of the sample mean difference changes under the alternative.

$$\bar{y}_{treat} - \bar{y}_{hist} \overset{H_1}{\sim} N(\delta^*, \sigma^2 * (\frac{1}{n_{treat}} + \frac{1}{n_{hist}}))$$

The effect size under the alternative δ^* must be subtracted before standardizing to arrive at a standard-normally distributed variable.

$$Z' = \frac{\bar{y}_{treat} - \bar{y}_{hist} - \delta^*}{\sigma\sqrt{\frac{1}{n_{treat}} + \frac{1}{n_{hist}}}} \overset{H_1}{\sim} N(0,1)$$

To arrive at the same test statistic as described above a non-centrality parameter ϑ must be added in a subsequent step.

$$T = \frac{Z' + \vartheta}{\sqrt{\frac{V}{df}}} = \left(\frac{\bar{y}_{treat} - \bar{y}_{hist} - \delta^*}{\sigma\sqrt{\frac{1}{n_{treat}} + \frac{1}{n_{hist}}}} + \frac{\delta^*}{\sigma\sqrt{\frac{1}{n_{treat}} + \frac{1}{n_{hist}}}}\right) * \sqrt{\frac{(N-2) * \sigma^2}{(N-2) * S^2}}$$

$$= \frac{\bar{y}_{treat} - \bar{y}_{hist}}{S_{pooled}\sqrt{\frac{1}{n_{treat}} + \frac{1}{n_{hist}}}} \qquad where \;\; \vartheta = \frac{\delta^*}{\sigma\sqrt{\frac{1}{n_{treat}} + \frac{1}{n_{hist}}}}$$

This results in the test statistic following a non-central t-distribution with $N-2$ degrees of freedom and ncp ϑ under the alternative (Johnson, Kotz and Balakrishnan, 1995, chap. 31):

$$T \overset{H_1}{\sim} t\left(df = N-2, ncp = \frac{\delta^*}{\sigma\sqrt{\frac{1}{n_{treat}} + \frac{1}{n_{hist}}}}\right)$$

Since the distribution of the test statistic under the alternative is dependent on sample size, probability of type-II-error and power are dependent on sample size as well. A sample size calculation is done to ensure the targeted power of a trial.

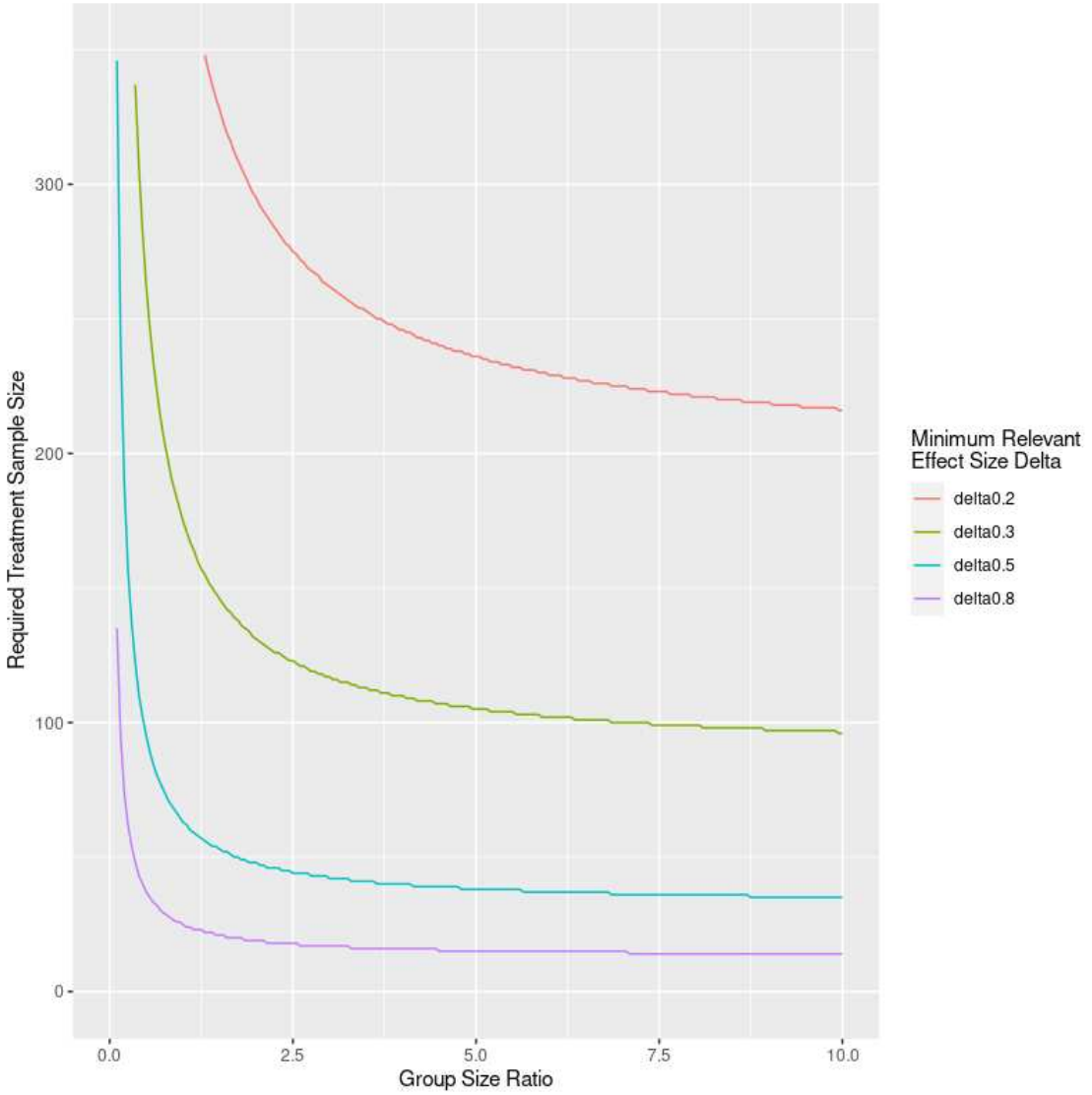


Figure 6: Approximate Sample Size over Group Size Ratio for the Two-Sample t-Test
This plot displays the approximate required treatment group sample size over different group size ratios. Different lines correspond to different alternative effect sizes "delta". A small alternative implies large sample size requirements. A large alternative implies smaller sample size requirements. Note that the approximate required control group size can is not displayed explicitly, but can be derived by multiplying the treatment group size by the corresponding group size ratio.

A closed form sample size formula for the two-sample t-test can be derived (Piantadosi, 2005, chap. 7) if quantiles of the normal distribution $Z_\alpha = \Phi^{-1}(1 - 0.025)$ and $Z_\beta = \Phi^{-1}(1 - 0.2)$ are used, where $\Phi^{-1}()$ denotes the quantile function of the

standard-normal distribution. An exact calculation using the quantiles of the t-distribution does not yield a closed sample size formula. The group size ratio is denoted by r.

$$n_{treat} \geq \frac{r+1}{r} * \frac{(Z_\alpha + Z_\beta)^2 \sigma^2}{\delta^{*2}}; \qquad r = \frac{n_{hist}}{n_{treat}}$$

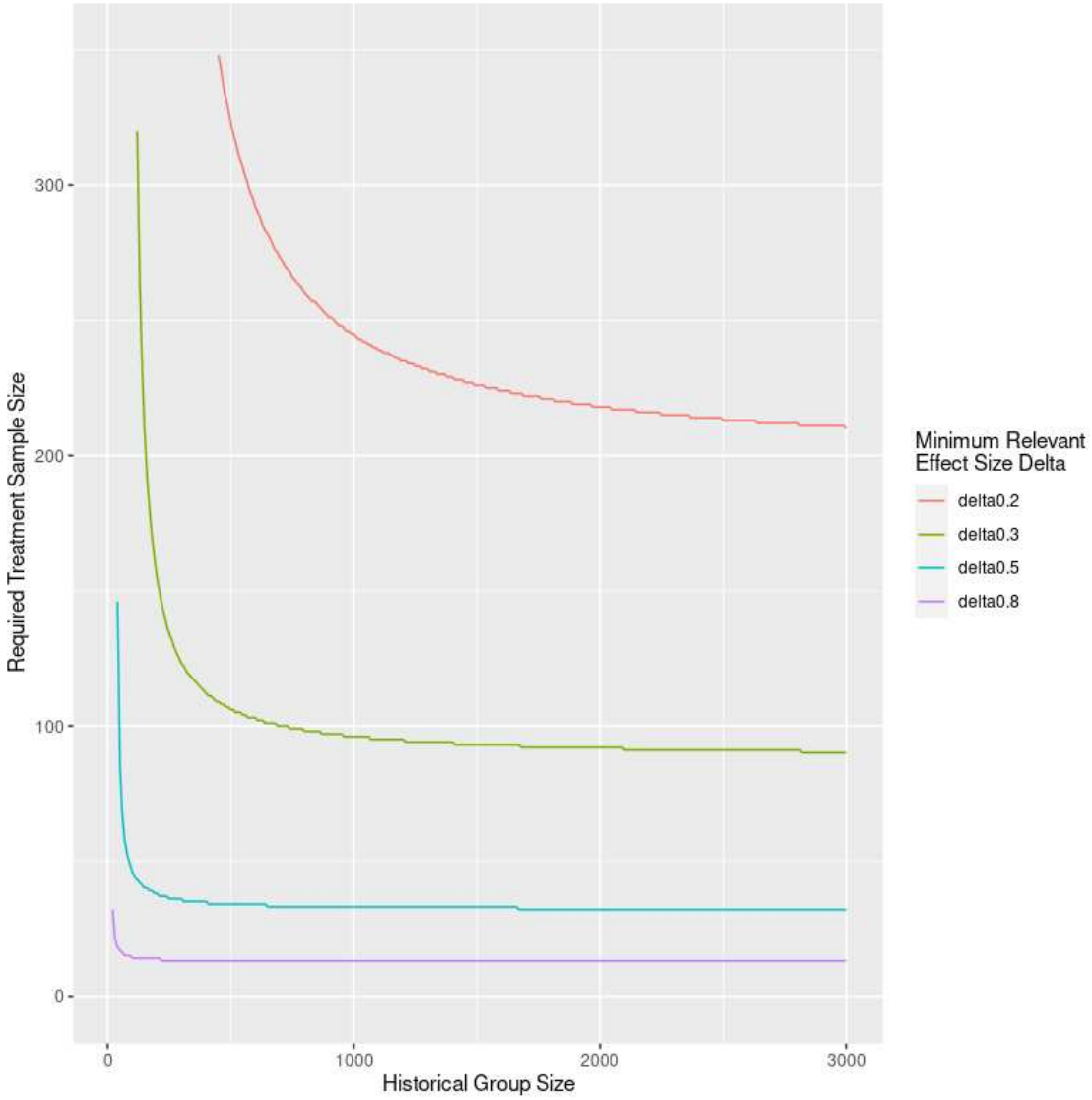


Figure 7: Approximate Sample Size over Historical Group Size for the Two-Sample t-Test
This plot displays the approximate required treatment group sample size over different historical group sizes. Different lines correspond to different alternative effect sizes "delta". A small alternative implies large sample size requirements. A large alternative implies smaller sample size requirements. Note that the group size ratio is not displayed explicitly, but can be derived by dividing the historical group size by the corresponding treatment group size.

Figure 6 displays the relation between group size ratio and approximate required treatment group size for 80% power, plotted for different alternatives.

In RCTs the group size ratio is under the influence of the trial designer. In contrast, in SATs a fixed size historical control cohort may be given, and treatment group size must be determined based on it. In this case it may be better to express the formula in terms of n_{hist}:

$$n_{treat} \geq \frac{1}{\frac{\delta^{*2}}{(Z_\alpha + Z_\beta)^2 \sigma^2} - \frac{1}{n_{hist}}}$$

Figure 7 displays the relation of the approximate required treatment group size and the historical sample size for 80% power, plotted for different alternatives.

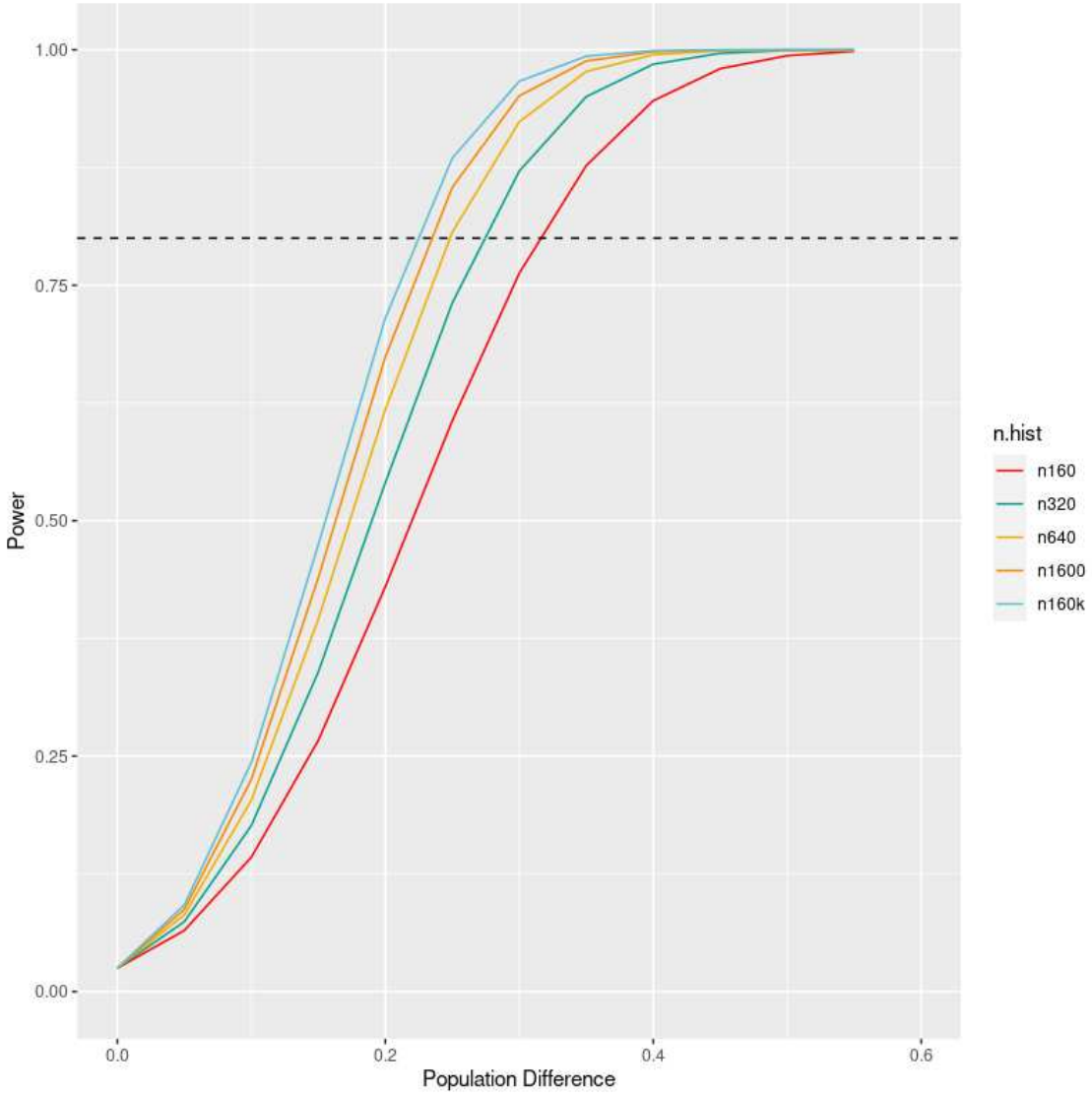


Figure 8: Power Curves for the Two-Sample t-Test for $n_{treat} = 160$ and $\sigma = 1$
The different power curves represent different group sizes used to conduct a two-sample t-test. The historical group size is given in the legend, while the treatment group size is equal for all curves. For very small population differences, power is close to zero in all settings. Power increases sharply for larger effect sizes if very large historical samples are used (blue curve). Power increases less rapidly if a relatively small historical group is used in the trial (red curve).

A power curve displays the power of a trial over the true effect size, that is the population difference as defined by the estimand. The power curves for the two-sample t-test in Figure 8 illustrate a property of designs with imbalanced group sizes. For the balanced case with both groups containing 160 patients, an effect size of 0.31 can be powered at 80%. For a historical control cohort of double this size, power gains are substantial, allowing an effect size of 0.28 to be powered. However, power gains become smaller if the external control group is further increased. A large external control group of 1600 patients permits powering of 0.24, while a very large historical control group of 160,000 patients permits an alternative of 0.23. Marginal power gains of additional external control units are *ceteris paribus* decreasing.

4.3 Rescaled T-Distribution

A relevant distribution in Section 5.3 of this thesis is the rescaled or non-standardized t-distribution. Note that it is different from the non-central t-distribution. Every probability distribution can be extended to a location-scale-family (Casella and Berger, 2002, chap. 3). A location-scale-family is a set of distributions, that only differ in expectation and variance. The normal distribution, for example, is immediately parametrized with location parameter μ, and scale parameter σ^2. The t-distribution, in contrast, is parametrized differently by degrees of freedom. Multiplying a t-distributed random variable T with a constant factor γ results in a variable following a rescaled t-distribution.

$$T * \gamma \overset{H_0}{\sim} t_{ls}(df = N - 2, scale = \gamma, loc = 0)$$

With probability density $f(t)$ and cumulative probability distribution $F(t)$ of the conventional (central) t-distribution, the density $d(x)$ and distribution $P(x)$ of the rescaled t-distribution can be calculated as follows (Casella and Berger, 2002, chap. 3):

$$d(x|\gamma) = \frac{1}{\gamma} f\left(\frac{T = t}{\gamma}\right); \qquad P(x|\gamma) = F\left(\frac{T = t}{\gamma}\right)$$

A rescaled t-distribution can arise if the two-sample t-test statistic is studentized by the standard error of the treatment mean: $\widehat{se}(\bar{y}_{treat}) = S_{treat}\sqrt{\frac{1}{n_{treat}}}$, instead of the standard error of the sample mean difference $\widehat{se}(\bar{y}_{treat} - \bar{y}_{hist}) = S_{pooled}\sqrt{\frac{1}{n_{treat}} + \frac{1}{n_{hist}}}$. Note that the former uses a different measurement variance estimate (S^2_{treat}) than the latter (S^2_{pooled}). S^2_{treat} is based on observations from the treatment group only, whereas

the pooled sample variance estimated S^2_{pooled} described above is based on both groups. Call this differently studentized test statistic Q. It is calculated as follows:

$$Q = \frac{\bar{y}_{treat} - \bar{y}_{hist}}{\widehat{se}(\bar{y}_{treat})} = \frac{(\bar{y}_{treat} - \bar{y}_{hist})\sqrt{n_{treat}}}{S_{treat}}$$

Numerator and denominator can be expanded by a scaling factor $\gamma = \sqrt{1 + \frac{n_{treat}}{n_{hist}}}$:

$$Q = \frac{\bar{y}_{treat} - \bar{y}_{hist}}{S_{treat}\sqrt{\frac{1}{n_{treat}}}\sqrt{1 + \frac{n_{treat}}{n_{hist}}}} * \sqrt{1 + \frac{n_{treat}}{n_{hist}}} = T * \gamma$$

The test statistic Q can be factored into two quantities: T and γ. The latter factor γ is a constant quantity, which acts as a scaling factor. The former factor T is a sample quantity, that is a random variable, and can be shown to be t-distributed with $n_{treat} - 1$ degrees of freedom. This is shown in the following by bringing T into the ratio form described in Section 244.2. Recap that the ratio of a standard-normally distributed variable and the root of a chi-square distributed variable divided by its degrees of freedom is t-distributed:

$$T = \frac{\bar{y}_{treat} - \bar{y}_{hist}}{S_{treat}\sqrt{\frac{1}{n_{treat}}}\sqrt{1 + \frac{n_{treat}}{n_{hist}}}} = \frac{\bar{y}_{treat} - \bar{y}_{hist}}{S_{treat}\sqrt{\frac{1}{n_{treat}} + \frac{1}{n_{hist}}}} = \frac{\bar{y}_{treat} - \bar{y}_{hist}}{\sigma\sqrt{\frac{1}{n_{treat}} + \frac{1}{n_{hist}}}} * \frac{\sigma}{S_{treat}}$$

$$= \frac{\bar{y}_{treat} - \bar{y}_{hist}}{\sigma\sqrt{\frac{1}{n_{treat}} + \frac{1}{n_{hist}}}} * \sqrt{\frac{(n_{treat} - 1) * \sigma^2}{(n_{treat} - 1) * S^2_{treat}}}$$

$$= \frac{Z}{\sqrt{\frac{V'}{n_{treat} - 1}}} \overset{H_0}{\sim} t(df = n_{treat} - 1)$$

This derivation is based on the assumption that S^2_{treat}, divided by the true variance σ^2 and multiplied by its degrees of freedom, is also chi-square distributed:

$$V' = (n_{treat} - 1)\frac{S^2_{treat}}{\sigma^2} \sim \chi^2(df = n_{treat} - 1)$$

What follows is that Q can be represented as a t-distributed variable T multiplied by a constant factor γ. Therefore Q follows a rescaled t-distribution with scaling factor γ:

$$Q \overset{H_0}{\sim} t_{ls}(df = n_{treat} - 1,\ scale = \gamma,\ loc = 0)$$

The finding that an incorrectly studentized sample mean difference $Q = \frac{\bar{y}_{treat} - \bar{y}_{hist}}{\widehat{se}(\bar{y}_{treat})}$ follows a rescaled t-distribution will be of central importance in Section 5.3.

4.4 Confounder Adjustment by MAIC

Confounding arises due to differences between trial arms with respect to baseline characteristics (Sterne *et al.*, 2016), as described in Section 2.2.1. To address confounding, baseline variables need to be modeled. A method to adjust for confounding, that is applicable in the case of restricted data availability, is MAIC. It was first suggested by Signorovitch *et al.* (2010) and extensively investigated by Cheng, Ayyagari and Signorovitch (2020). In the following a simple confounding scenario is outlined. The performance of MAIC in this scenario is investigated in Section 5.6.

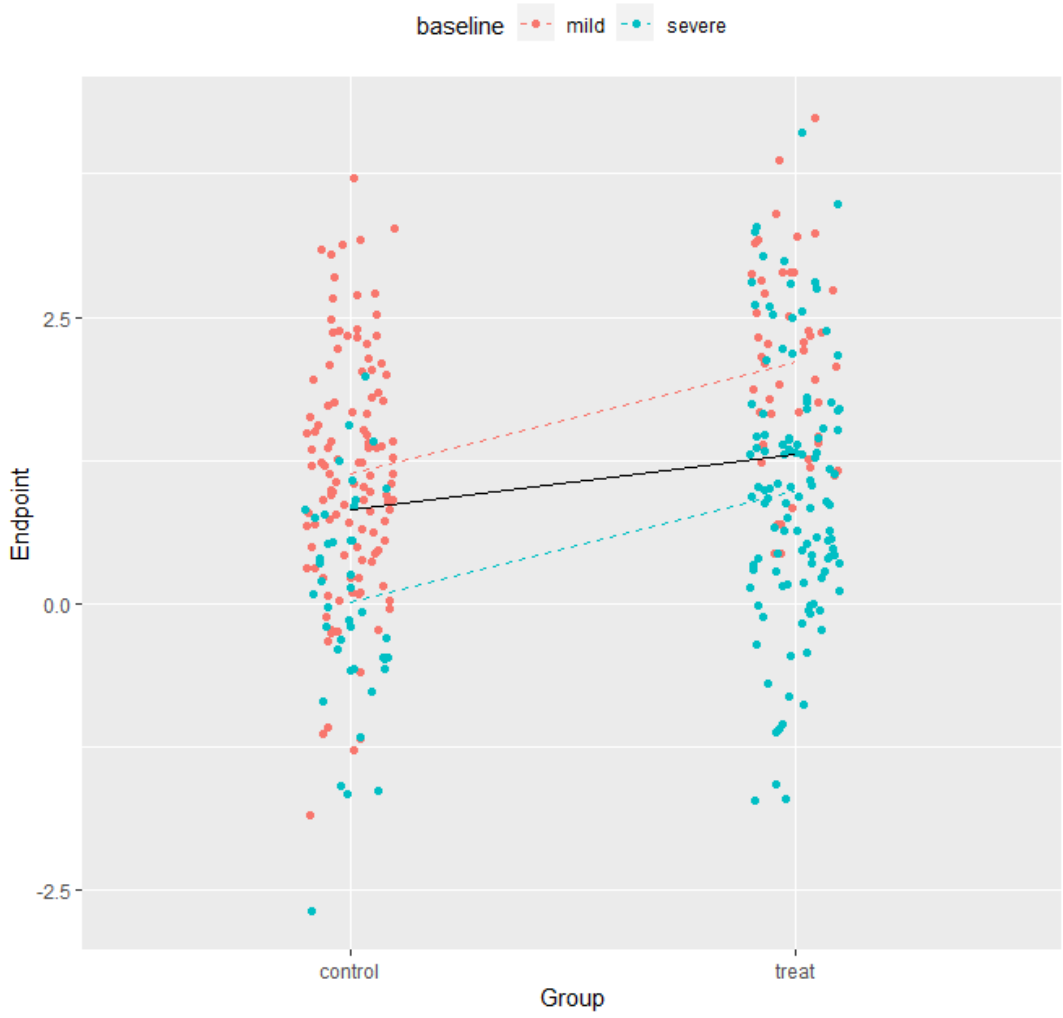


Figure 9: Confounding by a Binary Baseline, $\pi_{hist} = 0.75$, $\pi_{treat} = 0.25$
Red datapoints represent mildly diseased patients, blue datapoints represent severely diseased patients. The red stratum is larger in the control group, whereas the blue stratum is larger in the treatment group. Comparing corresponding stratum means in each group yields the colored dashed lines. These corresponds to the conditional treatment effect. Comparing the marginal means in each group yields the solid black line. It corresponds to the marginal treatment effect. A different slope results than for the conditional treatment effect.

Consider the case of a comparison between treatment and control arm, that is confounded by a binary variable “disease status at baseline” with factor levels “mild”

and “severe”. Let the baseline distribution of the confounder be different with respect to the trial arm: patients in the treatment arm are worse-off on average, while patients in the control arm are healthier on average. This is due to the higher share of mildly diseased patients in the control arm, while the treatment arm contains more severely diseased patients. The subgroups formed by mild and severe disease status are also referred to as *strata*.

In Figure 9 data is simulated and visualized under the assumption of equal treatment effects within strata. That is, a patient with severe disease status gets on average the same improvement of the endpoint as a patient with mild disease status. This is illustrated by the red and blue dashed lines displaying the same slope. If the naïve group means are compared, a smaller slope results, which is visualized by the solid black line.

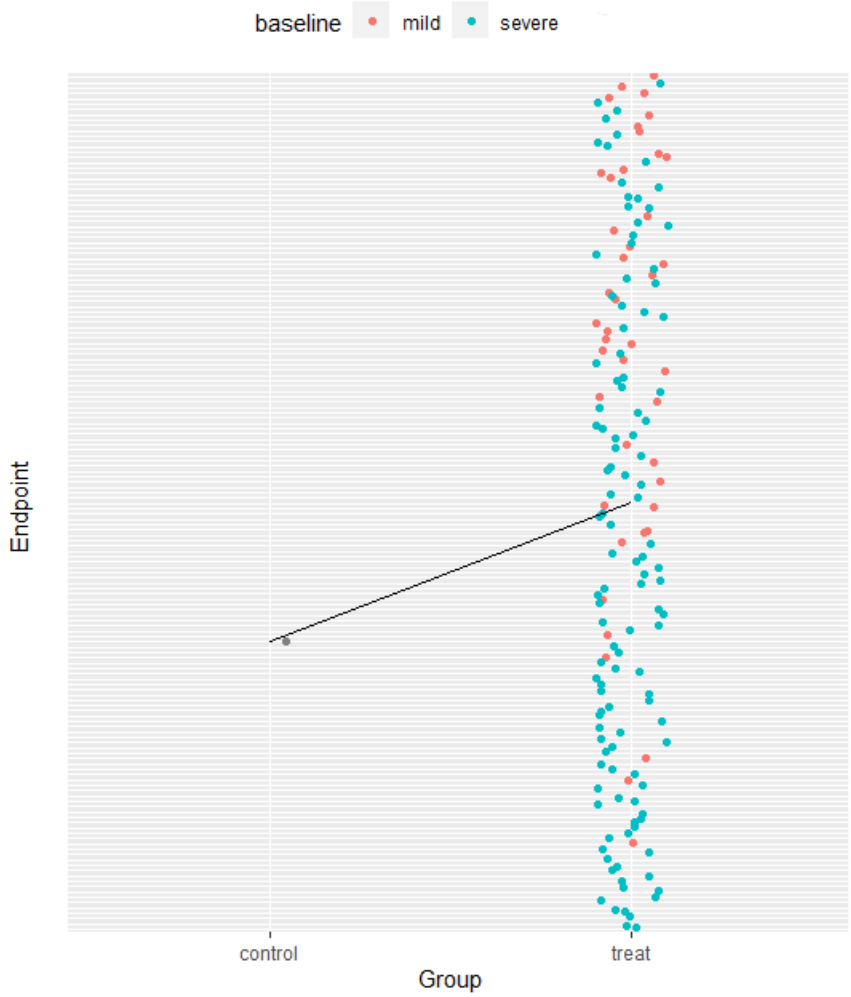


Figure 10: Confounding by a Binary Baseline - Restricted Data Availability - $\pi_{hist} = 0.75$, $\pi_{treat} = 0.25$. Information on the treatment group is available as IPD, while information on the control group is available as AGD. The marginal control mean is reported, which is represented by the black dot. The black line corresponds to the marginal treatment effect.

The naïve group means or marginal means are calculated equal as described in Section 4.2:

$$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{i,j}\,; \quad i = treat, hist;\ j = 1, \dots, n_i$$

The conditional stratum means are calculated using only datapoints in the respective subgroup:

$$\bar{y}_{i|k} = \frac{1}{n_{i,k}}\sum_{j=1}^{n_{i,k}} y_{i,k,j}\,; \quad k = mild, severe$$

The conditional treatment effect cannot be estimated unbiased if the marginal group means are compared. The marginal treatment effect underestimates the conditional treatment effect, that is of interest.

In case of restricted data availability on the external control group, the information may be visualized as displayed in Figure 10. A possibility to correct for confounding bias without IPD of the external control group is using MAIC. In the present case information on the historical control group consists in the marginal control mean $\bar{y}_{hist}$, as well an estimate of the relative stratum sizes within the historical control group $\hat{\pi}_{hist}$. The latter quantity is explained in detail below. Correction for confounding corresponds to an upwards correction of the marginal treatment mean in the treatment group, which is performed by a reweighting procedure, that is described in the following.

Call the binary confounder *"mild_status_base*" with coding "1" representing "mild", and "0" representing "severe". The distribution of disease status within a trial arm can be described by the Bernoulli distribution:

$$mild\_status\_base_{hist,j} \sim bern(\pi_{hist})$$
$$mild\_status\_base_{treat,j} \sim bern(\pi_{treat})$$
$$j = 1, \dots, n_i$$

Hence π_{hist} denotes the probability of a patient in the historical population being mildly diseases, while $1 - \pi_{hist}$ is the probability of a patient in the historical population being severely diseased. For the treatment the analogous consideration holds. Each trial arm is divided into two strata following the baseline parameters π_{hist} and π_{treat}, which are specific for the respective trial arm. The case of a single binary confounder is especially convenient, since the relative stratum sizes correspond to the estimates for the

baseline parameters. These are then used to estimate the treatment group mean in the historical population.

$$\widehat{\pi}_{hist} = \frac{n_{hist,mild}}{n_{hist}}; \quad \widehat{\pi}_{treat} = \frac{n_{treat,mild}}{n_{treat}}$$

$$\hat{\bar{y}}_{treat(hist)} = \widehat{\pi}_{hist} * \bar{y}_{treat|mild} + (1 - \widehat{\pi}_{hist}) * \bar{y}_{treat|severe}$$

Note that $\widehat{\pi}_{hist}$ is itself a sample quantity and therefore a random variable subject to sampling variation. The resulting distribution of the reweighted treatment mean $\hat{\bar{y}}_{treat(hist)}$ is a complex term. It can be simplified if stratum weights are assumed fix and known in advance:

$$C_1: \quad \widehat{\pi}_{hist} = \pi_{hist}$$

$$\bar{y}_{treat(hist)} = \pi_{hist} * \bar{y}_{treat|mild} + (1 - \pi_{hist}) * \bar{y}_{treat|severe}$$

The fix-weights-assumption is false in practice but may yield a good approximation if baseline parameters are not too close to the boundary values and group sizes are sufficiently large. A simulation study in Section 5.6 checks for robustness of this assumption.

The probability of a control patient having a mild-form-disease at baseline is higher than the corresponding probability in the control group: $\pi_{treat} < \pi_{hist}$. This imbalance is solved by upweighting the smaller treatment group stratum of patients with mild disease form on the one hand ($\pi_{hist} * \bar{y}_{treat|mild}$), while downweighing the larger stratum of severely diseased patients ($(1 - \pi_{hist}) * \bar{y}_{treat|severe}$).

5 Extending the Threshold-Crossing Analysis Framework

This section is structured as follows. Sections 5.1 introduces the TC framework as outlined in the proposing article of Eichler *et al.* (2016), while Section 0 describes the simulation study performed in the proposing article. Subsequently in Section 5.3, the central problem of this simulation study, which is coined *Variance-Adjustment Problem* (VAP) in this thesis, is explained and solved by using the rescaled t-distribution introduced in Section 4.3. In Section 5.4, the TC framework is extended to include heteroscedastic settings. In Section 5.5 the second problem of TC, which is coined *Bias-Adjustment Problem* (BAP) here, is discussed. In Section 5.6 the performance of MAIC as a solution to the BAP in a simple design as described in Section 4.4 is assessed.

5.1 Introduction to TC

TC is a framework for evidence generation in early phases of drug development programs. It aims for fast decisions in drug development programs, where large effects or no effects exist. The authors argue: “This is welcome because it is more important to provide timely access to highly beneficial than incremental treatments, and to terminate quickly and economically nonviable assets.” (Eichler *et al.*, 2016) This way TC is proposed to increase efficiency of pharmaceutical research. Its central feature is the use a SAT to assess the efficacy of the drug under investigation. External data is used as control arm. An outcome threshold is set based on the external control group information. If the treatment group outcome summary crosses the threshold, the drug may be judged effective. Otherwise, the drug is investigated in a further trial, which may be a RCT if practical or another SAT if not. Additionally, a futility threshold is set. If the treatment group outcome summary undercuts this futility threshold, the drug may be judged ineffective, and the drug development program is terminated. The TC framework as outlined in the proposing article by Eichler *et al.* (2016) is illustrated in Figure 11 and Figure 12.

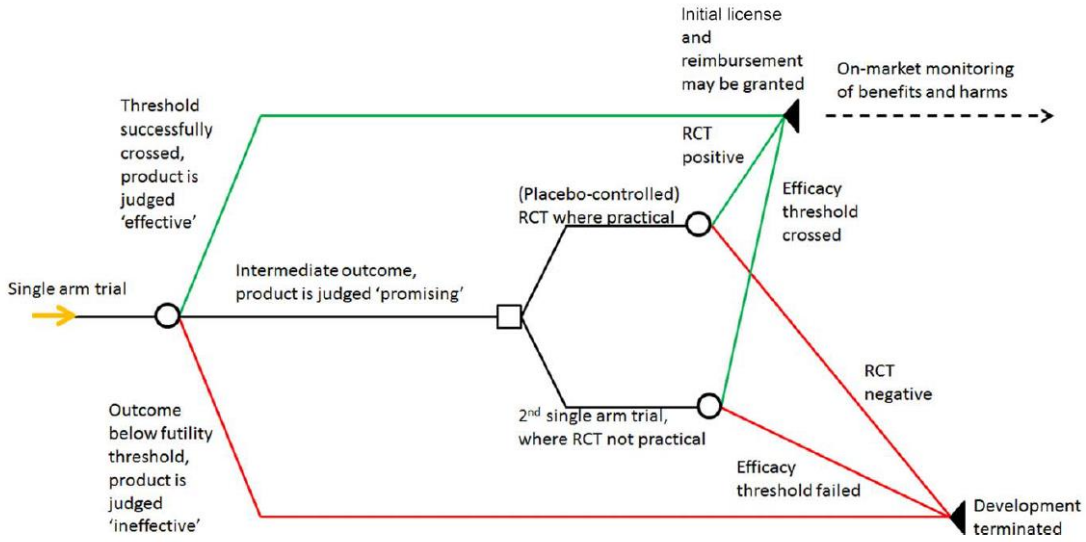


Figure 11: TC Framework - Drug Development Program (Eichler et al. 2016)
The decision framework starts after the conduct of an initial SAT (yellow arrow), as described in Figure 12. Drug effectiveness may be concluded if the prespecified threshold is crossed (green line). Similarly, ineffectiveness may be concluded if a futility threshold is undercut (red line). Inconclusive cases in between (black line) enter a second trial, which may be an RCT if applicable or a SAT if an RCT is inapplicable. The second SAT is evaluated by a second threshold. If crossed (green line), efficacy is concluded, if not crossed (red line) inefficacy is concluded.

Performing a trial with single arm design comes with a high potential for bias. However, the proposing article demands, that "even in the light of potential biases, the evidence-supporting efficacy of the drug should be unequivocal." (Eichler *et al.*, 2016) Sufficient certainty of results in the light of potentially biased effect estimates is possible in the case of dramatic effects (IQWiG, 2022a). The authors of the proposing article state that drugs with dramatic effects are rare, but they argue that they will be increasing in the upcoming years due to developments in cell-based and gene therapy, as well as personalized medicine (Eichler *et al.*, 2016). Additionally, applying RCT designs for investigating efficacy of drugs in these fields by is limited (Eichler *et al.*, 2016).

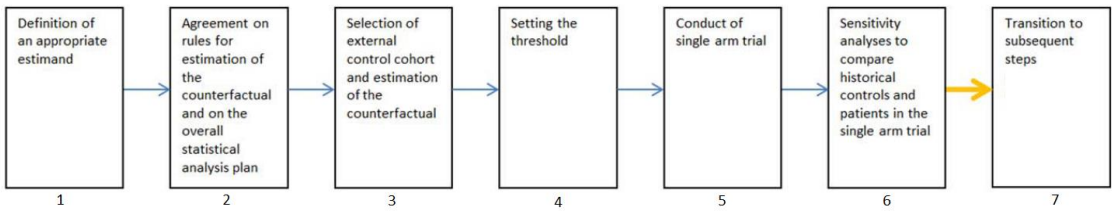


Figure 12: TC framework – Trial Conduct (Eichler et al. 2016)
This graph represents the step-by-step procedure for the trial conduct of the initial SAT in the TC framework. Blue arrows represent the transition to the next step. The yellow arrow represents the transition to the larger decision framework illustrated in Figure 11. Steps 1-4 constitute the prespecification of the estimand, the statistical analysis plan, the selection procedure for the external control group, as well as the threshold. The trial is conducted in Step 5. Sensitivity analyses follow in Step 6. The yellow arrow indicates the transition to subsequent steps, as outlined in Figure 11.

The conduct of the initial SAT is done in seven steps as illustrated in Figure 12. Steps 1 and 2 constitute the prespecification of the trial: Step 1 prespecifies the estimand, while Step 2 prespecifies the estimation, that is the statistical methods for analysis. In Step 3 the external control group is selected according to population criteria prespecified in Step 1. The estimation of the counterfactual corresponds to calculating the outcome summary measure of the control group. Based on this summary, the threshold is set in Step 4. The authors propose to perform this step in close coordination with authorities (Eichler *et al.*, 2016). Step 5 consists in the patient recruitment and treatment intervention. A potential crossing of the efficacy or futility threshold will manifest during this step. In Step 6 sensitivity analyses are performed to check for robustness of the methods used for in the main analysis. The implications of the trial for the drug development process are considered in Step 7. This step builds the bridge to the larger process of the drug development program illustrated in Figure 11.

5.2 Simulation Results of Eichler et al. (2016)

Eichler *et al.* (2016) use simulation studies to assess empirical type-I-error and power of a SAT. A simulation study is a computer experiment, where datapoints are sampled from a known data-generating process (Morris, White and Crowther, 2019). The statistical method under investigation is applied on these simulated datasets. The results of the method can be evaluated, since the true data-generating process is known. A single simulation run consists of an artificial dataset with group sizes denoted in the following as n_{treat} and n_{hist}. By conducting multiple simulation runs, the rates of correct and incorrect results of the statistical method can be evaluated. The number of total simulation runs will be denoted as n_{sim}. If data under the null hypothesis is generated, the rate of false rejection corresponds to the empirical type-I-error rate. In contrast, if data under the alternative is generated, the rate of correct rejections corresponds to the empirical power (Morris, White and Crowther, 2019).

In the investigated simulation designs on TC in Eichler *et al.* (2016), empirical type-I-error rate is not controlled. Either type-I-error inflation or -deflation is observed. The simulation designs consist of sampling two groups from a normal distribution (Eichler *et al.*, 2016). Observations scatter with equal measurement variance σ^2 around a group-specific mean μ_i.

$$y_{i,j} = \mu_i + \varepsilon_{i,j} \sim N(\mu_i, \sigma^2); \qquad i = treat, hist;\ j = 1, \dots, n_i$$

$$\varepsilon_{i,j} \sim N(0, \sigma^2)$$

Two simulation designs are investigated, called “no shift in time” and “shift in time”. The former design corresponds to an unbiased setting, the ladder to a biased one. The estimand δ is defined as the standardized difference between the true treatment mean μ_{treat} and the true counterfactual μ_{cf}, as would be measured by a parallel control group in an RCT.

$$\delta = \frac{\mu_{treat} - \mu_{cf}}{\sigma}$$

Unit-variance is assumed for convenience: $\sigma^2 = 1$. Empirical type-I-error rate and empirical power are investigated. For measuring the former, data is sampled under the null hypothesis, which corresponds to $\delta = 0$. For measuring empirical power, a small standardized effect size is set: $\delta = 0.2$. Note that group sizes n_i may differ in size, leading to a possibly unbalanced design.

Since an external control group is used only observations of μ_{hist} are available, while μ_{cf} cannot be assessed directly. In the unbiased design (“no shift in time”) the historical group can be used for unbiased estimation of the true counterfactual μ_{hist}. In the biased design (“shift in time”), external control observations systematically differ from the true counterfactual:

- Unbiased design: $\mu_{hist} = \mu_{cf}$
- Biased design: $\mu_{hist} + b = \mu_{cf}$

The hypothesis corresponding to the defined estimand is the following:

$$H_0: \quad \mu_{treat} - \mu_{cf} \leq 0$$

The significance level is set according to convention for one-sided tests as $\alpha = 2.5\%$. In Eichler *et al.* (2016), the conduct of this test is operationalized by setting a threshold thr and testing the treatment arm against it:

$$H_0^{thr}: \quad \mu_{treat} - thr \leq 0$$

For conducting the test, sample means $\bar{y}_i$ and sample variances S_i^2 and a threshold thr are required. Three different approaches of setting the threshold are discussed:

- Unadjusted: $thr_{unadj} = \bar{y}_{hist}$
- Variance-adjusted: $thr_{var.adj} = \bar{y}_{hist} + \frac{1}{2} CI_{hist}$
- Variance- & bias-adjusted: $thr_{double.adj} = \bar{y}_{hist} + \frac{1}{2} CI_{hist} + b * \delta$
    - With bias adjustment parameter $b \in \{0.1, 0.2, 0.3\}$

The calculation of the length of the historical confidence interval CI_{hist} will not be discussed here, since this approach is shown to be redundant in the following section.

The test statistic is not explicitly described in Eichler *et al.* (2016), but the authors presumably use the studentized difference of treatment group mean to threshold. The quantity used for studentizing presumably is the standard error of the treatment group mean. This test statistic is referred to as Q in the following.

$$Q = \frac{\bar{y}_{treat} - thr}{\widehat{se}(\bar{y}_{treat})}, \qquad \widehat{se}(\bar{y}_{treat}) = \frac{S_{treat}}{\sqrt{n_{treat}}}$$

The decision boundary Q^{crit} presumably is the $(1-\alpha)$-quantile of the t-distribution with $n_{treat} - 1$ degrees of freedom. These presumptions on the use of test statistic Q and decision boundary Q^{crit}, which are not explicitly described by Eichler *et al.* (2016), correspond to the conventional choices given a one-sample t-test of the treatment group against a constant value.

Empirical type-I-error rate is displayed for the two designs in the following graphs. Note that the different colored lines correspond to different approaches for setting the threshold. The unadjusted threshold thr_{unadj} is depicted in blue, the variance-adjusted threshold $thr_{var.adj}$ in black, and the double-adjusted thresholds $thr_{double.adj}$ in yellow, green and grey. The parallel group trial permits type-I-error control and is depicted in red.

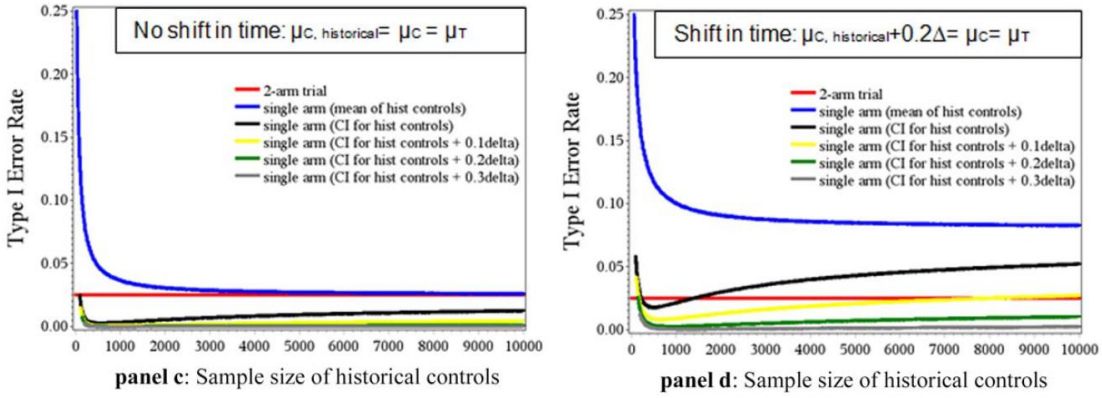


Figure 13: Simulation Results in Eichler et al. (2016)
In both plots type-I-error rate is depicted over different historical control group sizes. This is done for five different approaches of setting the threshold, as well as a conventional RCT design for benchmarking (red lines). In panel c (left-hand side) the results of simulating an unbiased design called "No shift in time" are depicted. In panel d (right-hand side) the results of simulating a biased design called "Shift in time" are depicted.

In the *unbiased design* depicted on the left side in Figure 13, using the unadjusted threshold leads to type-I-error inflation for small historical group sizes n_{hist}, as the blue line shows. For very small n_{hist} empirical type-I-error can be up to 25%. Call this the *Variance-Adjustment Problem* (VAP). Using the variance-adjusted threshold, type-I-error inflation can be averted, but only at the cost of type-I-error deflation, as the black line on the left side plot shows. The deflation is more pronounced for small n_{hist}. Type-I-error deflation comes with power loss and therefore higher sample size requirements (Eichler *et al.*, 2016).

In the *biased design* on the right-hand plot in Figure 13, type-I-error inflation occurs also if the variance-adjusted threshold (black line) is used. Call this the *Bias Adjustment Problem* (BAP). Using the bias- (and variance-) adjusted threshold permits keeping type-I-error, as the yellow, green and grey lines show. This again comes at the cost of type-I-error deflation and power loss. The effect is more pronounced the higher the bias adjustment parameter is chosen.

In summary, there are undesirable properties in the operating characteristics of the framework. Type-I-error is not controlled. Via adjusted threshold setting, empirical type-I-error is kept below significance level. In practice choosing the correct amount of bias adjustment will be difficult. Too little bias adjustment comes with type-I-error inflation, too much bias adjustment with type-I-error deflation and power loss.

5.3 Variance-Adjustment-Problem (VAP)

The VAP is the problem of type-I-error inflation in the *unbiased* simulation design, which is illustrated by the blue line on the left-hand side in Figure 13. Eichler *et al.* (2016) propose to solve it by raising the threshold by half the historical CI length.

$$thr_{var.adj} = \bar{y}_{hist} + \frac{1}{2} CI_{hist}$$

This comes at the cost of type-I-error deflation and power loss. In the following, the VAP is solved by analyzing properties of the test statistic Q used by Eichler *et al.* (2016). The solution implies that type-I-error probability can be controlled, type-I-error deflation and power losses can be averted.

If the threshold null hypothesis H_0^{thr} is desired to target the null hypothesis in line with the estimand $H_0: \mu_{treat} - \mu_{cf} \leq 0$, an auxiliary assumption A_1 is needed. This

assumption links the threshold to the historical population mean, and in the *unbiased design* ($\mu_{hist} = \mu_{cf}$) ultimately to the true counterfactual.

$$A_1: \quad thr_{unadj} = \mu_{hist}$$

Based on assumption A_1, the specification of the one-sample t-test is correct. The threshold is a constant value without variance and the distribution of the test statistic is:

$$Q = \frac{\bar{y}_{treat} - thr_{unadj}}{\widehat{se}(\bar{y}_{treat})} \overset{H_0, A_1}{\sim} t(df = n_{treat} - 1)$$

The quantile $Q^{crit} = t^{-1}(1 - \alpha,\ df = n_{treat} - 1)$ is used as a decision boundary and therefore type-I-error is controlled under the null and the auxiliary assumption:

$$P\left(Q > Q^{crit} \middle| H_0, A_1\right) = \alpha = 2.5\%$$

However, assumption A_1 does not hold. The unadjusted threshold, which is set as $thr_{unadj} = \bar{y}_{hist}$ does not equal μ_{hist}. Sampling variability will lead to random deviations of $\bar{y}_{hist}$ from μ_{hist}. The ignored variance in thr_{unadj} will be higher for small historical group sizes. This is what Eichler *et al.* (2016) observe in the unbiased simulation setting, shown on the left-hand side of Figure 13. Empirical type-I-error is high for low historical group sizes. For larger historical group sizes, empirical type-I-error rate converges towards the specified level. This reflects the fact that assumption A_1 is approximately true for large historical group sizes, while being violated considerably for small historical group sizes.

These empirical findings can be derived analytically by deriving the distribution of the test statistic in the absence of auxiliary assumption A_1. It can be shown that the test statistic Q follows a rescaled t-distribution with scale parameter γ, as introduced in Section 4.3.

$$Q \overset{H_0}{\sim} t_{ls}(df = n_{treat} - 1,\ scale = \gamma,\ loc = 0); \quad \gamma = \sqrt{1 + \frac{n_{treat}}{n_{hist}}}$$

The scaling factor γ induces stretching of the probability density of the test statistic Q. If the external control group size n_{hist} is large in comparison to the treatment group size, the influence of the scaling parameter γ is small since $\gamma \approx 1$. A plot of the density of Q for fixed treatment group size n_{treat} and different historical group size n_{hist} is given in Figure 14. It illustrates the reason of the observed type-I-error inflation: The stretching of the probability density function leads to a larger area under curve beyond the decision boundary Q^{crit} used in the one-sample t-test. This area corresponds to the

type-I-error probability. Lower historical group size leads *ceteris paribus* to a larger scaling factor and therefore a higher probability of type-I-error.

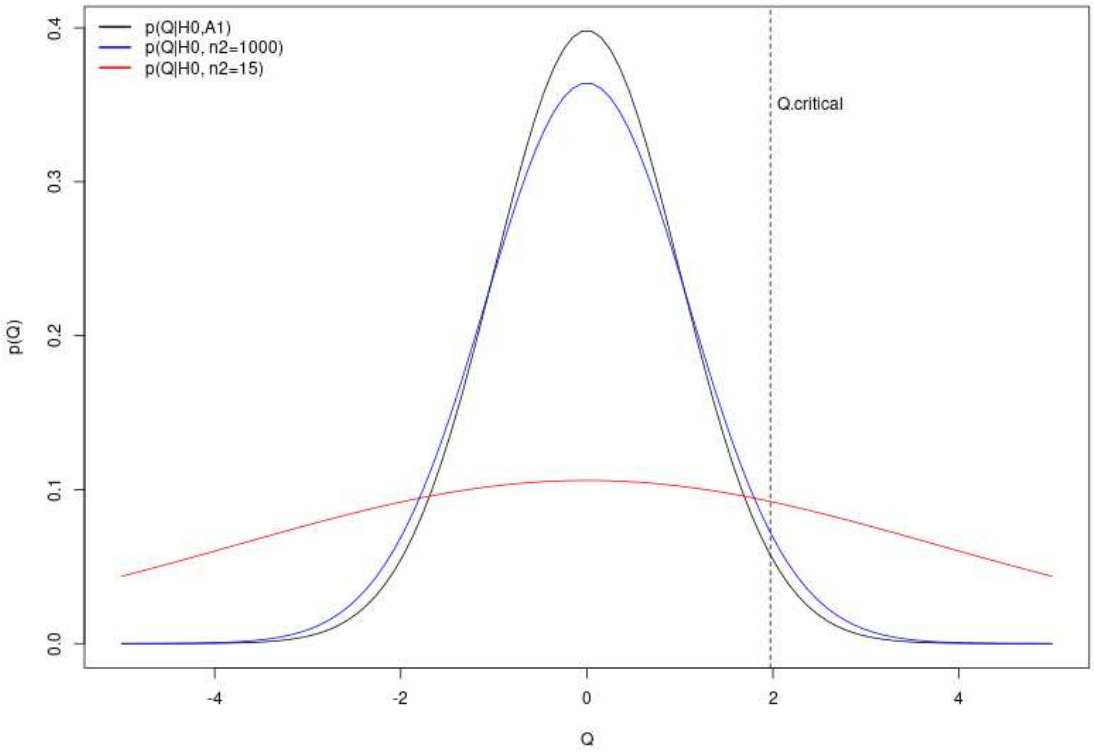


Figure 14: Probability Density of the Q-Statistic for $n_{treat} = 196$
The solid black line represents the probability density of the Q-Statistic under assumption A_1, which corresponds to a t-distribution. The dashed black line represents the critical value for the Q test, determined under assumption A_1. The colored lines represent the probability of the Q-Statistic if A_1 is not given, which correspond to rescaled t-distributions. For fix treatment group size a large historical group size induces only little stretching (blue line), whereas a small historical group size induces a large amount of stretching (red line). In both case the area under curve to the right of the critical value is larger compared to the distribution under A_1. This corresponds to an increased probability of type-I-error.

One solution to the VAP is using an adjusted decision boundary, that does not rely on auxiliary assumption A_1. Setting Q_{adj}^{crit} as the $(1-\alpha)$-quantile of the rescaled t-distribution results in a controlled type-I-error. It corresponds to the conventional t-quantile used by Eichler *et al.* (2016) multiplied by the scaling factor γ.

$$Q_{adj}^{crit} = {t_{ls}}^{-1}(1-\alpha,\ df = n_{treat} - 1,\ scale = \gamma,\ loc = 0) = \ \gamma * Q^{crit}$$

Call a one-sample t-test against the threshold with decision boundary adjustment as described above a "*q-test*". The validity of this q-test is double checked empirically by conducting a simulation study with data sampled under the null and tested by the q-test procedure (see Table 1).

Table 1: Empirical Type-I-Error Rate of the Q-Test for δ=0 – Homoscedastic Designs
Rejection rate of the null hypothesis across $n_{sim} = 100{,}000$ simulated datasets in each design is reported. Columns differ by testing method used for analysis of the dataset, as well as group size ratio r. Each dataset is sampled under the null hypothesis with $\delta = 0$, $\sigma = 1$ and $n_{treat} = 160$ group sizes following the group size ratio of the respective design and sample size formula given in this section. A simulation error of approximately 0.05 percentage-points results.

σ	*Group Size*	
	Balanced (r=1) n_{hist}*=160*	*Imbalanced (r=10)* n_{hist}*=1600*
1	2.54%	2.57%

For correct interpretation of simulation results, the simulation error or Monte Carlo error, which is due to the pseudo-random sampling of datasets, must be assessed (Morris, White and Crowther, 2019). The precision of the empirical error rates depends on the number of simulations runs n_{sim}. If few simulation runs are conducted, observed error rates may come with large variance. In this case, small deviations from a type-I-error target of 2.5% are not interpretable. The results of Table 1 come with a simulation error of approximately 0.05 percentage-points. This is calculated by the following formula (Morris, White and Crowther, 2019):

$$\widehat{se}(RejectionRate) = \sqrt{\frac{RejectionRate * (1 - RejectionRate)}{n_{sim}}}$$

Empirical type-I-error rates reported in Table 1 do not considerably differ from the target of 2.5%. The deviations by 0.04 and 0.07 percentage-points can be explained by simulation error.

A different solution to the VAP is using a two-sample t-test as introduced in Section 4.2. Here the formal testing procedure does not involve the threshold thr_{unadj}, but the historical group mean $\bar{y}_{hist}$. Statistically, this makes no difference, since the former is set as equal to the latter in the q-test: $thr_{unadj} = \bar{y}_{hist}$. On a practical level, however, some differences arise, which are discussed in the conclusion of the thesis (Section 6). The null hypothesis is equal to the q-test:

$$H_0: \quad \mu_{treat} - \mu_{cf} \leq 0$$

The conventional two-sample t-test statistic is used. Its distribution is known and probability of type-I-error is controlled.

$$T = \frac{\bar{y}_{treat} - \bar{y}_{hist}}{S_{pooled}\sqrt{\frac{1}{n_{treat}} + \frac{1}{n_{hist}}}} \overset{H_0}{\sim} t(1-\alpha, df = N-2)$$

One may expect the two-sample t-test to perform better in terms of power than the one-sample q-test with decision boundary adjustment due to the fact, that information of both groups is used in the test statistic. However, an empirical investigation by means of a simulation study shows hardly any differences between the methods (Table 2). Sample size calculation for both methods is performed via the normal approximation of the two-sample t-test introduced in Section 4.2:

$$n_{treat} \geq \frac{r+1}{r} * \frac{\left(Z_\alpha + Z_\beta\right)^2 \sigma^2}{\delta^{*2}}; \qquad n_{hist} \geq r * n_{treat}$$

Table 2: Empirical Power for $\delta = 0.3$ – Homoscedastic Designs
Rejection rate of the null hypothesis across $n_{sim} = 100{,}000$ simulated datasets in each design is reported. Columns differ by testing method used for analysis of the dataset, as well as group size ratio r. Each dataset is sampled under the alternative hypothesis with $\delta = 0.3$, $\sigma = 1$ and group sizes following the group size ratio of the respective design and sample size formula given in this section. A simulation error of approximately 0.1 percentage-points results.

	Testing Procedure			
	Two-Sample t-test		One-sample "Q"-test	
	Group Size			
σ	*Balanced (r=1)*	*Imbalanced (r=10)*	*Balanced (r=1)*	*Imbalanced (r=10)*
1	79.8% (n_{treat}=175)	79.9% (n_{treat}=96)	79.8% (n_{treat}=175)	79.2% (n_{treat}=96)

In case of balanced group sizes empirical power is equal. If group sizes are strongly unbalanced the two-sample t-test slightly outperforms the one-sample q-test.

5.4 Heteroscedastic Setting – Welch-test

In the last subsection a solution to the VAP is proposed, that permits type-I-error control analytically in the unbiased design. In this section the unbiased design is extended by the heteroscedastic setting.

Testing for a group mean difference when measurement variance is not assumed equal across groups with possibly unequal group sizes is known as the Behrens-Fisher problem (Brunner, Bathke and Konietschke, 2018, chap. 3). This situation is especially important for SATs, since group sizes are expected to be unequal, and the assumption

of equal variance is questionable in case different data sources are used. Measurements in both groups are normally distributed and scatter around the respective group mean μ_i with group-specific variance ${\sigma_i}^2$:

$$y_{i,j} \sim N(\mu_i, {\sigma_i}^2); \quad i = treat, hist; \; j = 1, \dots, n_i; \quad \sigma_{hist} = \rho * \sigma_{treat}$$

The t-test is known to be flawed in a Behrens-Fisher design. The case of negative pairing and positive pairing are distinguished. In case of negative pairing the larger group possesses smaller variance. If the larger group possesses larger variance the setting is referred to as positive pairing. The t-test performs liberal in case of negative pairing. This means that type-I-error rate is inflated, while type-II-error rate is decreased. An increase in power results, which is considered artificial. It is merely the consequence of type-I-error inflation. Conservative performance of the t-test is present in case of positive pairing. This implies decreased type-I-error rate, as well as decreased power (Brunner, Bathke and Konietschke, 2018, chap. 3).

The definition of the causal effect of interest is the first problem to encounter in the Behrens-Fisher problem. Using the standardized effect difference raises the question of which quantity to use for standardization. Since there is better information on the treatment group, due to the active patient recruitment, the treatment group variance is chosen here:

$$\delta = \frac{\mu_{treat} - \mu_{hist}}{\sigma_{treat}}$$

The null hypothesis is as follows:

$$H_0: \mu_{treat} - \mu_{cf} \leq 0$$

It is tested using the Welch-test, which is considered a good approximate solution to the Behrens-Fisher problem (Brunner, Bathke and Konietschke, 2018, chap. 3). Note that type-I-error control in the Welch-test is not shown by statistical theory as it is for the t-test. Performance assessment therefore relies on empirical investigations by simulation studies. The test statistic of the Welch-test is as follows:

$$W = \frac{\bar{y}_{treat} - thr_{unadj}}{\sqrt{\frac{S^2_{treat}}{n_{treat}} + \frac{S^2_{hist}}{n_{hist}}}}$$

It uses separate variance estimates for each group, instead of a pooled estimate. The distribution of the test statistic W cannot be attained in closed analytical form. The Welch-test approximates it using a t-distribution, whose degrees of freedoms are calculated via the Welch-Satterthwaite equation:

$$\hat{v} = \frac{(\frac{S_{treat}^2}{n_{treat}} + \frac{S_{hist}^2}{n_{hist}})^2}{\frac{S_{treat}^4}{{n_{treat}}^2(n_{treat}-1)} + \frac{S_{hist}^4}{{n_{hist}}^2(n_{hist}-1)}}$$

$$W \overset{H_0, approx.}{\sim} t(df = \hat{v})$$

The decision boundary is $W^{crit} = t^{-1}(q = 1-\alpha,\ df = \hat{v})$. The approximative nature of the test prohibits calculation of theoretical type-I-error rate. However, empirical type-I-error rate can be checked in a simulation study. Table 3 reports empirical type-I-error rate of the Welch-test. The central row represents the homoscedastic setting as a special case. Type-I-error rates are in close proximity to the target of 2.5%, with differences from observed rate to target explainable by the simulation error of 0.05 percentage-points. This is also true for rows 2 and 4, representing heteroscedastic settings, as well as for the first row, representing an extreme heteroscedastic setting with a ratio of standard deviations $\rho = \frac{\sigma_{hist}}{\sigma_{treat}} = 0.1$, resulting in possibly negative pairing. In the last row, representing extreme heteroscedasticity with possibly positive pairing, empirical type-I-error rate is off the 2.5%-target by two simulation errors. A slight type-I-error deflation in the balanced case is indicated, while inflation is indicated for the unbalanced case observed. These extreme settings are included to check thoroughly the Welch-test's performance. In practice an extreme variance ratio would raise doubt about the overall comparability between the trial arms.

Table 3: Empirical Type-I-Error Rate - Welch-Test
Rejection rate of the null hypothesis across $n_{sim} = 100{,}000$ simulated datasets in each design is reported. Columns differ by group size ratio r. With fixed treatment group size of n_{hist}=160 in all designs, historical group size is fixed for each column as displayed in the column header. Each dataset is sampled under the null hypothesis with $\delta = 0$, $\sigma_{treat} = 1$. Rows of the table represent varied standard deviations σ_{hist} of historical measurements. A simulation error of approximately 0.05 percentage-points results.

σ_{hist}	*Group Size*	
	Balanced (r=1) n_{hist}=160	*Imbalanced (r=10)* n_{hist}=1600
0.1	2.55%	2.48%
0.5	2.45%	2.44%
1	2.54%	2.53%
2	2.55%	2.49%
10	2.38%	2.60%

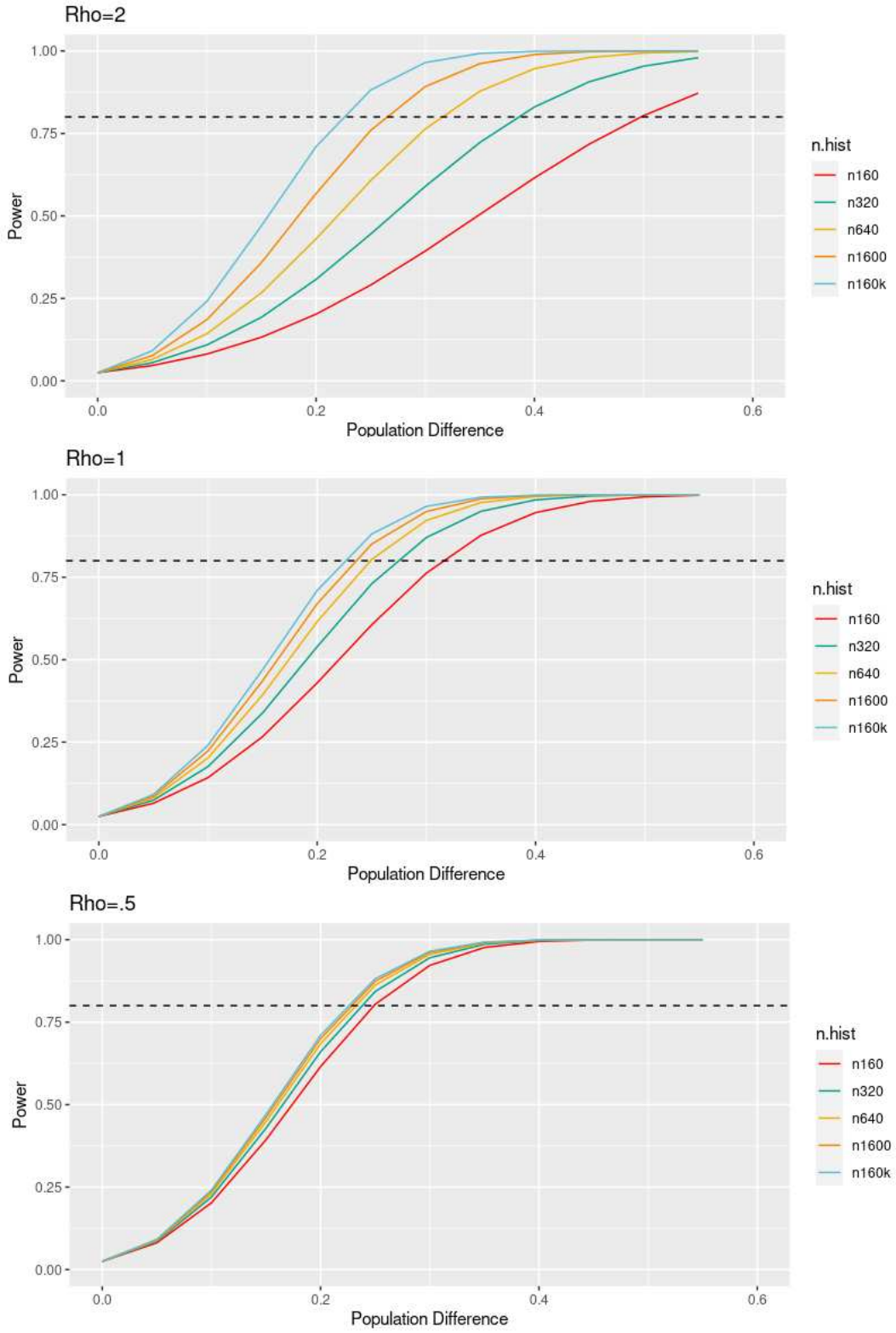


Figure 15: Power Curves - Welch Test for $n_{treat} = 160$ and $\sigma_{treat} = 1$
Upper panel: positive pairing. Mid panel: homoscedastic design. Lower panel: negative pairing. The different power curves represent different group sizes used to conduct a Welch-test. The historical group size is given in the legend, while the treatment group size is equal for all curves. For very small population differences, power is close to zero in all settings. Power increases sharply for larger effect sizes if very large historical samples are used (blue curves). Power increases less rapidly if a relatively small historical group is used in the trial (red curves).

The power curves in Figure 15 are displayed for the homoscedastic setting where $\rho = 1$ (mid panel), and the moderately heteroscedastic settings with positive pairing with $\rho = 2$ (upper panel) and negative pairing with $\rho = 0.5$ (lower panel). The curves in the homoscedastic setting are equivalent to the power curves for the two-sample t-test displayed in Figure 8 (Section 4.2). In the positive pairing design, power is lower, while in the negative pairing design it is higher compared to the homoscedastic setting. Note that power curves are equal in all designs for very large historical group sizes used (blue curve).

For checking empirical power, an approximate sample size calculation is performed in each simulation design according to the following formula, which is a simplified version of (Schouten, 1999). Variance ratio is parametrized by $\tau = \rho^2$:

$$n_{treat} \geq \frac{(r+\tau)}{r} * \frac{\left(Z_\alpha + Z_\beta\right)^2 * \sigma_{treat}^2}{\delta^{*2}}$$

For fix historical group size, the following formula can be used:

$$n_{treat} \geq \frac{1}{\frac{\delta^{*2}}{\left(Z_\alpha + Z_\beta\right)^2 \sigma_{treat}^2} - \frac{\tau}{n_{hist}}}$$

Table 4: Empirical Power of the Welch-Test for $\delta = 0.3$ – Heteroscedastic Designs
Rejection rate of the null hypothesis across $n_{sim} = 100{,}000$ simulated datasets in each design is reported. Columns represent a balanced and an unbalanced design. Each dataset is sampled under the alternative hypothesis with $\delta = 0.3$, $\sigma = 1$ and group sizes following the group size ratio of the respective design and sample size formula given in this section. A simulation error of approximately 0.1 percentage-points results.

	Group Size	
σ_{hist}	*Balanced (r=1)*	*Imbalanced (r=10)*
0.1	79.6% (n_{treat}=89)	79.6% (n_{treat}=89)
0.5	80.0% (n_{treat}=110)	79.7% (n_{treat}=90)
1	80.0% (n_{treat}=175)	79.4% (n_{treat}=96)
2	80.0% (n_{treat}=437)	79.9% (n_{treat}=123)
10	80.2% (n_{treat}=8809)	80.1% (n_{treat}=960)

Table 4 shows empirical power for the Welch-test, as well as group sizes used in each design. The mid row shows group sizes needed in the homoscedastic setting. In the balanced design group sizes of 175 patients are needed while in the unbalanced case

96 patients for the treatment group are needed and the external control consists of 960 patients. The resulting power in the balanced case is similar up to simulation error to the two-sample t-test, while in the imbalanced case the Welch-test displays slightly less power. If historical variance is raised, higher group sizes are needed to power the trial. If historical variance is lowered, smaller group sizes are sufficient for an empirical power of almost 80%.

5.5 Bias-Adjustment-Problem (BAP)

The VAP is a question of *statistical precision*. Eichler et al. (2016) found type-I-error inflation in the unbiased design, which was caused by sampling variability. This problem was shown to be solvable by choosing the correct statistical analysis. Type-I-error can be controlled in the homoscedastic setting, while in the heteroscedastic setting simulation studies confirm that type-I-error rate is sufficiently close to the significance level.

The BAP denotes the phenomenon of type-I-error inflation, which is due to systematic differences between trial arms. Therefore, it is a question of *internal validity* of the trial. Eichler et al. (2016) discuss a biased design, which is called “shift in time”. Here the observed historical control group is systematically different than the true counterfactual, as would be measured by a randomized parallel control group: $\mu_{hist} \neq \mu_{cf}$. A motivation for the naming as time-shift bias is an improving standard of care. This makes the historical control patients systematically worse off than hypothetical parallel control patients collected at the time of the trial: $\mu_{hist} < \mu_{cf}$.

Comparing the treatment arm to the historical control arm yields an overestimated treatment effect. If a hypothesis test is performed, type-I-error inflation occurs. Keeping type-I-error under control in the biased design is not a question of correct statistical modelling, as was the case for the VAP. In principle a quantitative adjustment to the counterfactual can adjust for the bias and keep type-I-error at the specified level. Eichler et al. (2016) suggest raising the threshold by some amount. In practice the quantification of the correct amount for adjustment will be difficult. The correct amount is the difference between historical control mean and true counterfactual $b = \mu_{cf} - \mu_{hist}$. However, information on the true counterfactual μ_{cf} is not measured, so there is no empirical basis to estimate b. The difference in standard of care between historical

and parallel control patients is hard to quantify. If the analyst does an attempt for quantification there is no means to check for resulting validity, so the question of the correct adjustment remains ex-post. Choosing an incorrect bias adjustment comes with disadvantages. Too little adjustment yields a liberal hypothesis test, type-I-error is still inflated, albeit less than in the unadjusted case. Too much of bias adjustment leads to a conservative hypothesis test with deflated type-I-error and little power.

A scenario more interesting for statistical modeling is measured bias in the form of confounding. In the following the MAIC-model containing a binary baseline confounder is investigated.

5.6 MAIC to address the BAP

MAIC demands targeting the trial arm, for which only AGD is available, which in practice will be the external control arm. The definition of the treatment effect includes the counterfactual treatment group mean in the historical population $\mu_{treat(hist)}$. It is estimated by the reweighting procedure described in Section 4.4:

$$\bar{y}_{treat(hist)} = \pi_{hist} * \bar{y}_{treat|mild} + (1 - \pi_{hist}) * \bar{y}_{treat|severe}$$

The estimand is defined as the standardized population mean difference in the historical control population. The measurement variance of the treatment arm σ_{treat} is chosen for standardizing.

$$\delta = \frac{\mu_{treat(hist)} - \mu_{hist(hist)}}{\sigma_{treat}}$$

The threshold is set as the marginal outcome mean in the historical group: $thr = \bar{y}_{hist}$. The decision on efficacy is based on a hypothesis test with null hypothesis as follows.

$$H_0: \quad \mu_{treat(hist)} - thr = 0$$

A Welch-type test statistic is used. This allows for different measurement variance between trial arms. The numerator of the test statistic consists of the sample mean difference using the reweighted treatment sample mean: $\bar{y}_{treat(hist)} - \bar{y}_{hist(hist)}$. Studentizing this term requires the sample variances of the marginal historical mean $\widehat{var}(\bar{y}_{hist}) = S^2_{hist} \frac{1}{n_{hist}}$ and of the reweighted treatment sample mean $\widehat{var}\left(\bar{y}_{treat(hist)}\right) = \lambda * S^2_{treat} \frac{1}{n_{treat}}$. The derivation of the latter term follows in the next paragraph, leading to the product of the marginal treatment group variance multiplied by an adjustment factor λ. This result in the following test statistic, which is approximately t-distributed under

the null hypothesis with degrees of freedom estimated by the Welch-Satterthwaite approximation:

$$M = \frac{\bar{y}_{treat(hist)} - thr_{unadj}}{\sqrt{\lambda * \frac{S^2_{treat}}{n_{treat}} + \frac{S^2_{hist}}{n_{hist}}}} \overset{H_0, approx.}{\sim} t(df = \hat{v})$$

$$\hat{v} = \frac{(\frac{S^2_{treat} * \lambda}{n_{treat}} + \frac{S^2_{hist}}{n_{hist}})^2}{\frac{S^4_{treat} * \lambda^2}{n^2_{treat}(n_{treat} - 1)} + \frac{S^4_{hist}}{n^2_{hist}(n_{hist} - 1)}}$$

The variance adjustment factor λ is determined by considering the variance of the reweighted treatment sample mean under the fix-weights-assumptions:

$$C_1: \quad \hat{\pi}_{hist} = \pi_{hist}$$

$$C_2: \quad \hat{\pi}_{treat} = \pi_{treat}$$

Variances of the conditional means in the treatment group are as follows:

$$var(\bar{y}_{treat|k}) = \sigma^2_{treat} \frac{1}{n_{treat,k}}; \; k = mild, severe$$

Conditional means in the treatment group can be considered independent:

$$C_3: \quad cov(\bar{y}_{treat,mild}, \bar{y}_{treat,severe}) = 0$$

Recap, that the relation between stratum sizes and the baseline parameters are as follows:

$$n_{treat,mild} = \hat{\pi}_{treat} * n_{treat}; \qquad n_{treat,severe} = (1 - \hat{\pi}_{treat}) * n_{treat}$$

Given these assumptions the following term can be derived:

$$\begin{aligned} var(\bar{y}_{treat(hist)}) &= var(\pi_{hist} * \bar{y}_{treat,mild} + (1 - \pi_{hist}) * \bar{y}_{treat,severe}) \\ &= \pi^2_{hist} * var(\bar{y}_{treat,mild}) + (1 - \pi_{hist})^2 * var(\bar{y}_{treat,severe}) \\ &\quad + 2 * \pi_{hist} * (1 - \pi_{hist}) * cov(\bar{y}_{treat,mild}, \bar{y}_{treat,severe}) \\ &= \pi^2_{hist} * \sigma^2_{treat} \frac{1}{n_{treat,mild}} + (1 - \pi_{hist})^2 * \sigma^2_{treat} \frac{1}{n_{treat,severe}} + 0 \\ &= \pi^2_{hist} * \frac{\sigma^2_{treat}}{n_{treat} * \pi_{treat}} + (1 - \pi_{hist})^2 * \frac{\sigma^2_{treat}}{n_{treat} * (1 - \pi_{treat})}) \\ &= \sigma^2_{treat} \frac{1}{n_{treat}} * \left(\frac{\pi^2_{hist}}{\pi_{treat}} + \frac{(1 - \pi_{hist})^2}{1 - \pi_{treat}} \right) \\ &= \sigma^2_{treat} \frac{1}{n_{treat}} * \lambda, \qquad where \; \lambda = \frac{\pi^2_{hist}}{\pi_{treat}} + \frac{(1 - \pi_{hist})^2}{1 - \pi_{treat}} \end{aligned}$$

The relation of the baseline parameters and the adjustment factor λ is as follows and illustrated in the following plot:

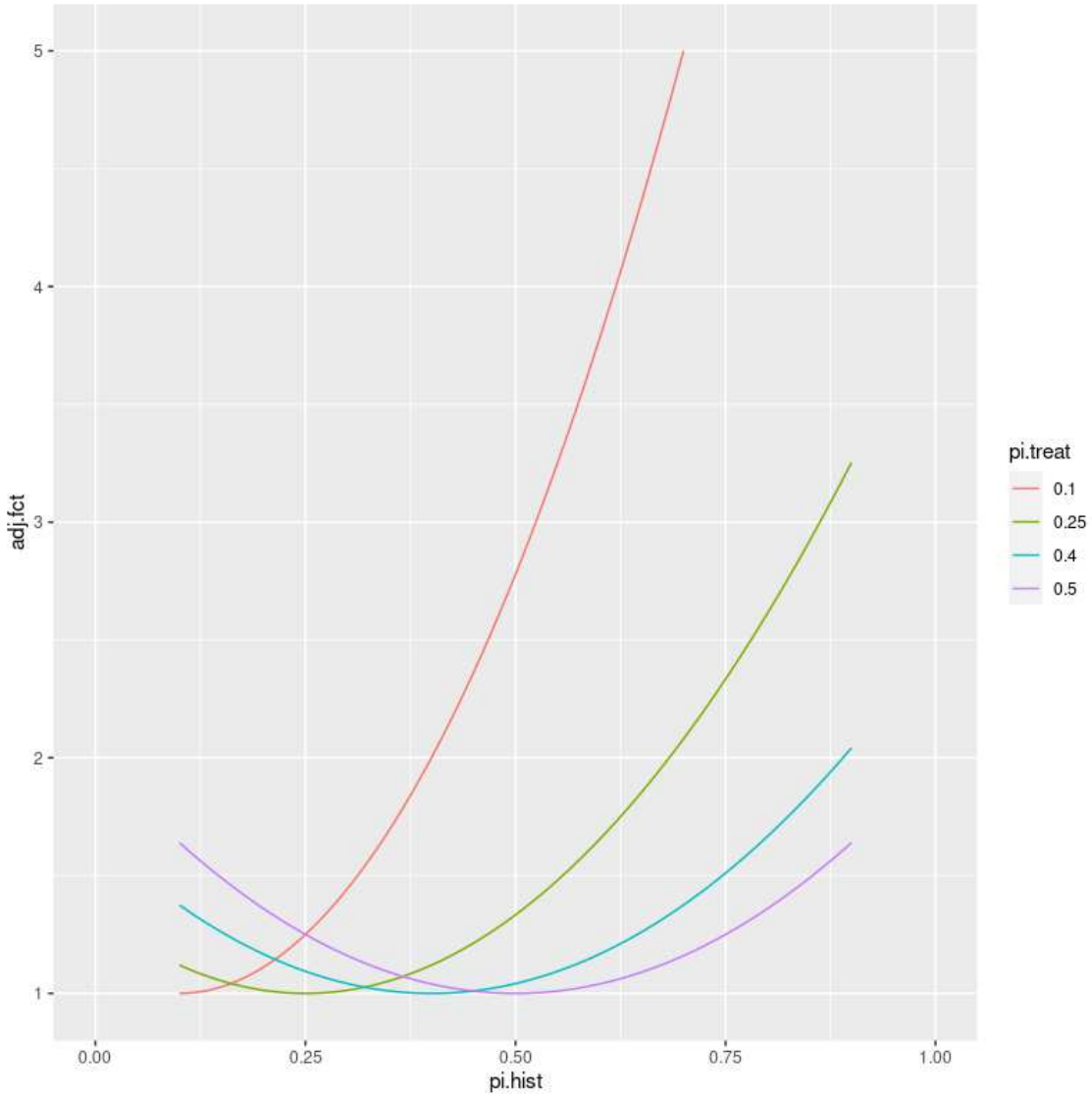


Figure 16: Variance Adjustment Factor λ over historical baseline parameter π_{hist}, factored by treatment baseline parameter π_{treat}. If baseline parameters match, the adjustment factor equals 1. The factor increases parabolically with increasing baseline imbalance between groups.

In case of balance in baseline parameters between trial arms $\pi_{treat} = \pi_{hist}$, the adjustment factor λ equals 1. In this case no reweighing takes place, the marginal treatment mean can be used for an unbiased comparison.

If baseline parameters differ, the adjustment factor λ is greater than 1. This implies a higher variance of the reweighted treatment mean $\widehat{var}\left(\bar{y}_{treat(hist)}\right)$. Moderate differences in baseline parameters imply a moderate adjustment factor of $\lambda \leq 2$ (see

blue or purple curve). When parameters are close to the boundary values, λ can become very large. Note that higher estimate-variance implies higher sample size requirements, which will be explained in detail further below.

A simulation study is conducted to check if type-I-error rate stays close to the nominal level of 2.5%. Empirical baseline estimates $\hat{\pi}_{treat}$ and $\hat{\pi}_{hist}$ are used for reweighting the treatment mean and for estimating variance:

$$\hat{\bar{y}}_{treat(hist)} = \hat{\pi}_{hist} * \bar{y}_{treat|mild} + (1 - \hat{\pi}_{hist}) * \bar{y}_{treat|severe}$$

$$\widehat{var}\left(\hat{\bar{y}}_{treat(hist)}\right) = {S_{treat}}^2 \frac{1}{n_{treat}} * \hat{\lambda}$$

Data under the null hypothesis of no treatment effect is generated.

$$H_0: \mu_{treat|k} - \mu_{hist|k} = 0;\ k = mild, severe$$

Constant design parameters are:

- Baseline difference $\Delta = \mu_{treat|mild} - \mu_{treat|severe} = \mu_{hist|mild} - \mu_{hist|severe} = 0.3$
- Measurement variance of the treatment arm $\sigma_{treat} = 1$
- Treatment group size $n_{treat} = 160$
- Historical baseline parameter $\pi_{treat} = 0.75$

The following design parameters are set to varied levels:

- Measurement variance of the historical arm $\sigma_{hist} \in \{0.1, 0.5, 1, 2, 10\}$
- Historical group size $n_{hist} \in \{160, 1600\}$
    - This translates to group size ratios of $r \in \{1, 10\}$
- Treatment group baseline parameter $\pi_{treat} \in \{0.75, 0.5, 0.25\}$
    - This translates into an unconfounded, a moderately confounded and a highly confounded design

Crossing these design factors yields unique 12 designs. For each design $n_{sim} = 100.000$ simulation runs are performed, and the empirical rate of null rejection is measured.

Empirical type-I-error in Table 5 is below nominal level of 2.5% in all designs. Most designs show moderate deflation with type-I-error between 2.1 and 2.5%. The deflation is more pronounced for designs with negative pairing, that is low historical variance. Group size imbalance and baseline imbalance do not seem to have influence on empirical type-I-error.

To check for the influence of the baseline difference Δ between stratum means within the groups, a second simulation study with design parameter $\Delta = 1$ is performed. Its results are displayed in Table 6. The type-I-error deflation pattern is more pronounces here.

Table 5: Empirical Type-I-Error – MAIC-adjusted Welch-Test - Δ=0.3
Rejection rate of the null hypothesis across $n_{sim} = 100{,}000$ simulated datasets in each design is reported. Columns differ by baseline imbalance, as well as group size ratio r. Each dataset is sampled under the null hypothesis with $\delta = 0$, $\sigma_{treat} = 1$ and $n_{treat} = 160$ and historical group size n_{hist} as displayed in the respective column. A simulation error of approximately 0.05 percentage-points results.

	Baseline Imbalance			
	Moderate: π_{treat}=0.5		Large: π_{treat}=0.25	
	Group Size			
σ_{hist}	*Balanced (r=1)* n_{hist}=160	*Imbalanced (r=10)* n_{hist}=1600	*Balanced (r=1)* n_{hist}=160	*Imbalanced (r=10)* n_{hist}=1600
0.1	2.22%	2.27%	2.14%	2.17%
0.5	2.26%	2.19%	2.13%	2.13%
1	2.28%	2.29%	2.29%	2.19%
2	2.41%	2.38%	2.39%	2.18%
10	2.52%	2.44%	2.52%	2.47%

Table 6: Empirical Type-I-Error – MAIC-adjusted Welch-Test - Δ=1
Rejection rate of the null hypothesis across $n_{sim} = 100{,}000$ simulated datasets in each design is reported. Columns differ by baseline imbalance, as well as group size ratio r. Each dataset is sampled under the null hypothesis with $\delta = 0$, $\sigma_{treat} = 1$ and $n_{treat} = 160$ and historical group size n_{hist} as displayed in the respective column. A simulation error of approximately 0.05 percentage-points results.

	Baseline Imbalance			
	Moderate: π_{treat}=0.5		Large: π_{treat}=0.25	
	Group Size			
σ_{hist}	*Balanced (r=1)* n_{hist}=160	*Imbalanced (r=10)* n_{hist}=1600	*Balanced (r=1)* n_{hist}=160	*Imbalanced (r=10)* n_{hist}=1600
0.1	0.85%	1.17%	0.99%	1.19%
0.5	1.04%	1.11%	1.10%	1.14%
1	1.35%	1.27%	1.39%	1.20%
2	1.94%	1.43%	1.86%	1.30%
10	2.50%	2.28%	2.49%	2.24%

If reweighting takes place, the concept of effective sample size (ESS) is of interest. Some patients are upweighted and some downweighed in the MAIC-procedure. This comes at the cost of not having full information of the treatment group entering the mean estimation and hypothesis test. ESS can be calculated based on the individual

weights according to a formula given in the online appendix of Signorovitch *et al.* (2010):

$$ESS = \frac{(\sum_{i=1}^{n_{treat}} \omega_i)^2}{\sum_{i=1}^{n_{treat}} {\omega_i}^2}$$

Here ω_i denotes the individual weight received by treatment group patient i, where the weights are in standardized form, that is adding up to 1. Until this point only the aggregate stratum weights π_{hist} and $(1 - \pi_{hist})$ were considered. These corresponded to the historical baseline parameter and are convenient to derive properties of the reweighted mean estimate. In contrast, individual weights ω_i are based on the ratio between historical and concurrent baseline parameters:

$$\omega_i = \frac{\pi_{hist}}{\pi_{treat}} * \frac{1}{n_{treat}} \quad , mildly\ diseased\ patients$$

$$\omega_i = \frac{1 - \pi_{hist}}{1 - \pi_{treat}} * \frac{1}{n_{treat}} \quad , severely\ diseased\ patients$$

The choice of individual weights as the baseline-ratios above is based on its equivalence to the stratum-weights form introduced in Section 4.4:

$$\bar{y}_{treat(hist)} = \pi_{hist} * \bar{y}_{treat|mild} + (1 - \pi_{hist}) * \bar{y}_{treat|severe}$$

For a derivation of this equivalence, see the appendix of this thesis. Using these individual weights results in the aggregate weights being as specified above. Based on the fix-weights-assumptions C_1 and C_2, the ESS formula equals the following form:

$$ESS = \frac{(\sum_{i=1}^{n_{treat}} \omega_i)^2}{\sum_{i=1}^{n_{treat}} {\omega_i}^2} = \frac{1}{\lambda} * n_{treat}$$

For a detailed derivation see as well the appendix.

The approximate sample size formula used in the empirical power investigation is the following:

$$n_{treat} \geq \frac{\lambda r + \tau}{r} * \frac{\left(Z_\alpha + Z_\beta\right)^2 \sigma_{treat}^2}{{\delta^*}^2}; \quad n_{hist} \geq r * n_{treat}$$

If required treatment group size is expressed in terms of a fixed historical sample size, the difference to the Welch-test sample size formula is the factor λ.

$$n_{treat} \geq \frac{\lambda}{\frac{{\delta^*}^2}{\left(Z_\alpha + Z_\beta\right)^2} - \frac{\tau}{n_{hist}}} = \lambda * n_{treat,undj}$$

The MAIC-reweighting procedure to address confounding bias comes at the cost of higher sample size demands compared to the unadjusted analysis by a Welch-test. The higher sample size demand is dependent on the amount of reweighting of the

treatment group, that is needed for adjustment. It can be parametrized conveniently by λ, given the model assumptions.

For power analysis a standardized effect of $\delta^* = 0.3$ is chosen.

$$H_1: \mu_{treat|k} - \mu_{hist|k} = \delta^*; \qquad k = mild, severe$$

Under the alternative hypothesis the test statistic approximately follows a non-central t-distribution with degrees of freedom $\hat{v}$, estimated by the Welch-Satterthwaite equation and ncp $\hat{\vartheta}$ dependent on variance estimates, including the variance adjustment factor λ:

$$\hat{\vartheta} = \frac{\delta^*}{\sqrt{\lambda * \frac{S_{treat}}{n_{treat}} + \frac{S_{hist}}{n_{hist}}}}$$

$$M \overset{H_1, approx.}{\sim} t(df = \hat{v}, ncp = \hat{\vartheta})$$

The decision boundary $M^{crit} = t^{-1}(q = 1 - \alpha,\ df = \hat{v}, ncp = 0)$ remains unchanged. Power corresponds to the are under the probability density beyond the critical value given the alternative is true $P_{H_1}\left(M > M^{crit}\right) = 1 - T(M^{crit}, df = \hat{v}, ncp = \vartheta)$. Calculation of power curves is dependent on the same assumptions done as in the theoretical investigation of type-I-error probability:

- Fix weights assumptions C_1, C_2 to calculate the expected λ
- Independence of stratum means within the treatment group
    - $cov\left(\bar{y}_{treat,mild}, \bar{y}_{treat,severe}\right) = 0$

Figure 17 shows three power curves. For the red line group sizes are chosen to achieve 80% power for a standardized effect of $\delta^* = 0.3$. Measurement variance is assumed equal. The yellow and green curve illustrate an interesting finding: increasing the treatment group size comes with higher power gains, than increasing the external control group size by the same amount. This finding can be explained by the reweighting procedure, that applies only to the treatment group.

The empirical power check (Table 7) depends neither on the fix-weights-assumptions C_1 and C_2, nor on the zero-covariance-assumption C_3. Empirical baseline estimates are used for reweighting, empirical variance estimates used for studentizing the test statistic. Data under the alternative hypothesis $\delta = 0.3$ is sampled and a hypothesis

test based on the MAIC-reweighting procedure is performed. Designs are equal to the ones used for investigating empirical type-I-error.

Higher sample sizes are needed compared to the Welch-test. The sample size formula derived is a sufficient approximation to achieve close to 80% power. Power is closest to 80% in the designs with very high historical variance.

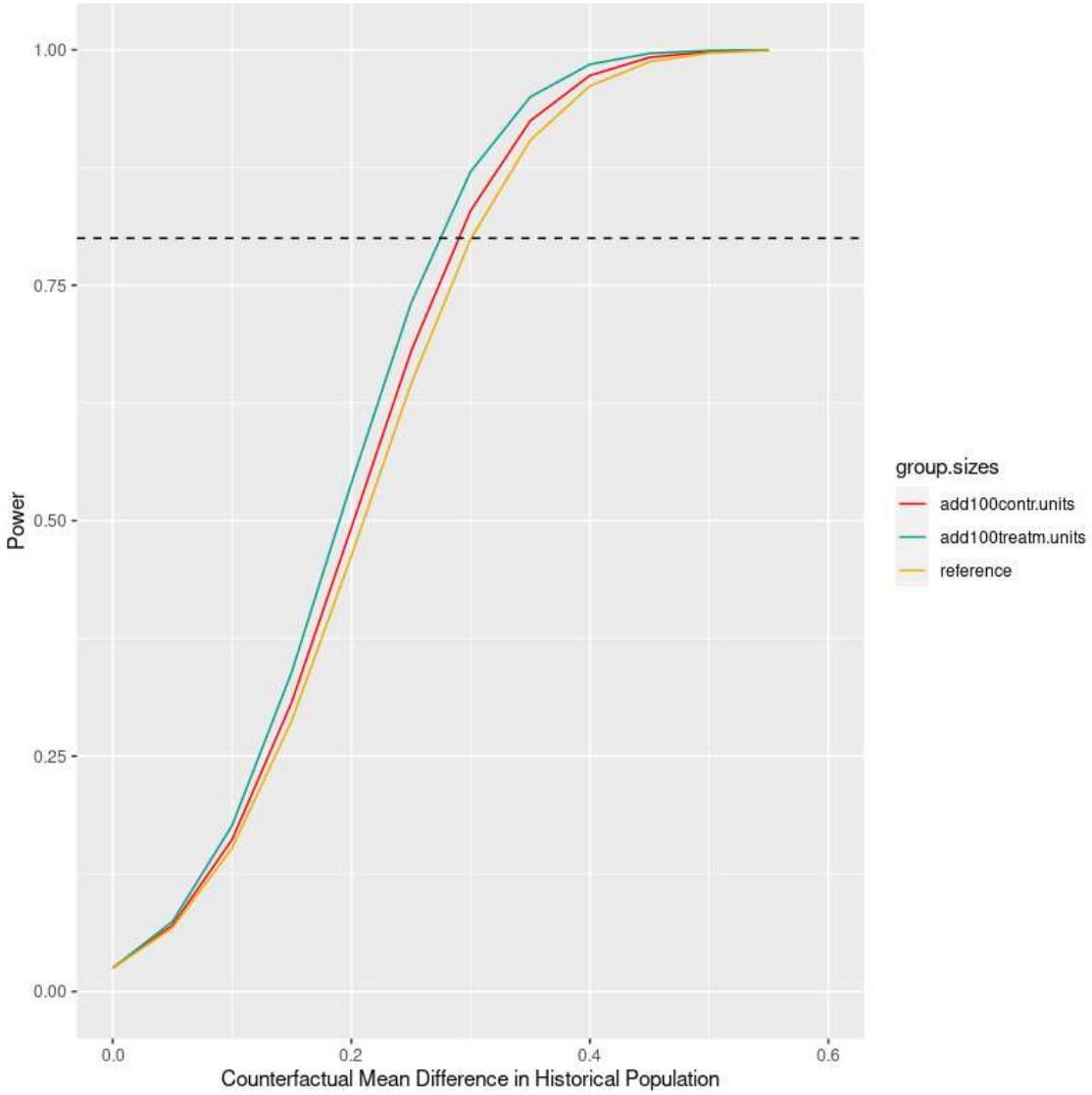


Figure 17: Power Curves – MAIC, $\sigma_{treat} = 1$, $\sigma_{hist} = 1$
The yellow curve displays power over varied values of the estimand, defined as the counterfactual mean difference in the historical population. For this reference curve, group sizes are set to $n_{treat} = 291$ and $n_{hist} = 291$, which yields 80% power for an effect of $\delta^* = 0.3$ standard deviations. Adding 100 additional treatment units (green curve) increases power more than adding 100 additional control units (red curve).

Table 7: Empirical Power for δ=0.3 – MAIC-adjusted Welch-Test
Rejection rate of the null hypothesis across $n_{sim} = 100{,}000$ simulated datasets in each design is reported. Columns represent a balanced and an unbalanced design. Each dataset is sampled under the alternative hypothesis with $\delta = 0.3$, $\sigma_{treat} = 1$ and group sizes following the group size ratio of the respective design. A simulation error of approximately 0.1 percentage-points results.

	Baseline Imbalance			
	Moderate: *π_{treat}=0.5*		Large*: π_{treat}=0.25*	
	Group Size			
σ_{hist}	*Balanced (r=1)*	*Imbalanced (r=10)*	*Balanced (r=1)*	*Imbalanced (r=10)*
0.1	78.0% (n_{treat}=110)	79.1% (n_{treat}=110)	78.9% (n_{treat}=205)	79.2% (n_{treat}=204)
0.5	78.7% (n_{treat}=131)	79.0% (n_{treat}=112)	79.2% (n_{treat}=226)	79.3% (n_{treat}=206)
1	79.4% (n_{treat}=197)	78.9% (n_{treat}=118)	79.3% (n_{treat}=291)	79.3% (n_{treat}=213)
2	79.7% (n_{treat}=458)	79.2% (n_{treat}=144)	79.7% (n_{treat}=553)	79.3% (n_{treat}=239)
10	79.9% (n_{treat}=8830)	79.9% (n_{treat}=982)	80.1% (n_{treat}=8925)	79.9% (n_{treat}=1076)

6 Conclusions and Discussion

6.1 Threshold-Crossing

The relevance of the TC framework at the present day is questionable. Eichler *et al.* (2016) already point out that dramatic effects make a small share of drugs investigated in pharmaceutical research. Review studies confirm this claim (Ribeiro *et al.*, 2022). Arguments are invoked that novel bio-technology and personalized medicine increase the occurrence of dramatic effects (Eichler *et al.*, 2016). However, making a "prediction is very difficult, especially if it's about the future"[1].

Assumed dramatic effects increase in the future, it is questionable if TC contributes to increasing efficiency in pharmaceutical research. On the one hand dramatic effects are found more rapid and with less costs. On the other hand, inconclusive cases must be investigated as RCT at a later stage in the drug development process. Hence two trials are performed, where a single RCT would have been sufficient. This reduces efficiency of pharmaceutical research.

In some drug development programs RCTs may be infeasible. This is the case in rare diseases. Here legal arrangements, such as the Orphan drug status, already exist to increase incentives of manufacturers to invest into drug development. The Ide-Cel case (Section 3) exemplifies a benefit decision based on the Orphan drug status. SAT evidence had been submitted for benefit quantification, but was rejected due to bias concerns and the absence of a dramatic effect size (G-BA, 2022b). Nonetheless a non-quantifiable benefit was granted, which is an important step towards reimbursement of the treatment by the German health system and ultimately may justify the investment of the manufacturer into drug development. TC would have brought no advantage in the drug development program of Ide-Cel. The comparisons conducted for benefit assessment were at high risk of bias. This would have demanded a high bias adjustment of the threshold. Since no dramatic effect was present the threshold would presumably not have been crossed. Hence the investigation of the drug would have been required to be repeated within a second trial. In the Ide-Cel case, applying TC would presumably decrease efficiency.

[1] The origin of the quote is unclear, while most quoters attribute it to physicist Niels Bohr.

If a SAT is applied for evidence generation, strategies exist to reduce the potential for bias. An important aspect is prespecification of trial conduct. This includes prespecification of how to select an external control cohort. (Eichler *et al.*, 2016) state that “cherry-picking” an external control cohort yields a potential for bias.
Setting an “a priori fixed [outcome] threshold” constitutes an important part of prespecification, according to Eichler *et al.* (2016). As explained in Section 5.3, this led the authors to choose the incorrect statistical method for analysis. In the simple homoscedastic and unbiased design described in Section 5.3, an adjustment of the decision boundary permits the setting of a fixed threshold. In Sections 5.4 and 5.6 of this thesis designs are investigated, where this is not straight-forward possible. Setting a fixed threshold rather implies inconvenience in the statistical analysis. If TC is to be applied in practice, more complex designs will be required. Setting a threshold in these designs presumably results in statistical imprecision, that cannot be mitigated by a simple adjustment of the decision boundary as seen in Section 5.3. However, these difficulties are not of major concern. The important aspect of TC is not the setting of a fix threshold to compare the treatment group against. Importance should be on prespecification of trial conduct. Acceptability for authorities and public trust does not depend on the setting of a fix threshold as a constant, but on the credibility of compliance to the prespecified statistical analysis. Incorrect methods for statistical analysis should in any case be avoided.

Considering bias adjustment, TC seemingly allows for a convenient means, that is raising the threshold. A higher threshold, in turn, demands higher effect sizes present, in order to be detected. In the case of dramatic effects this is an applicable strategy. However, if the treatment effect is small and a raised threshold is applied, the comparison against the threshold may result in a negative effect. If this leads also to the undercutting of a futility threshold, the drug development program is stopped, even if a small treatment effect was present. If not, the trial is inconclusive and further trials are conducted.

The preferable choice to reduce bias is by design. If the bias source is measured it is can be adjusted for in the statistical analysis. Selection bias is minimized by matching population criteria of the cohorts, confounding bias can be addressed by statistical methods, such as MAIC. In practice both tasks will be limited by data availability of the external control cohorts. As discussed in the example of Ide-Cel, data availability

prohibited the assessment of selection and confounding bias. For benefit assessment of Ide-Cel, the treatment cohort of the approval-trial was indirectly compared to external control cohorts. To reduce data availability problem a close communication between conductors of the approval trial and conductors of benefit assessment trials is required. Especially if past RCT cohorts are used as controls, baseline information tends to be sparser than in non-randomized designs, since confounding is of less concern in RCTs. A systematic research of relevant confounders may be especially useful before setting up patient registries for specific diseases or indications. Collecting vast information on possible confounders allows the comparability to SAT cohorts in the future.

6.2 External Validity

In Section 2.7, the concept of external validity is mentioned. External validity refers to the transferability of trial results to everyday clinical practice. Some sources of external data, especially transactional data, are considered to reflect better clinical practice than data collected within a trial. An example is the “healthy-volunteer effect” (Ford and Norrie, 2016), which occurs if study participants are systematically better-off than patients in usual care. Using transactional data or as control group in a SAT is argued to increase external validity of the trial. However, if this comes at the cost of internal validity, nothing is gained (Hoffmann *et al.*, 2021). If the trial result is internally invalid, the invalid effect estimate cannot be generalized to standard clinical care settings. Ford and Norrie remark that the distinction between internal and external validity is too simplistic (Ford and Norrie, 2016). They promote the conduct of “large simple trials” or *pragmatic trials* (Ford and Norrie, 2016), which are characterized by more flexibility in recruitment of patients, follow-up and other categories. Pragmatic trials demand that participants be “similar to patients who would receive the intervention if it became usual care.” (Ford and Norrie, 2016). The reduced effort spent on questionnaires and follow-up aims to increase trial participation. Still ensuring high-quality trials is of high priority. Ford and Norrie suggest to decide on implementing pragmatic features based on the individual trial objective.

References

Austin, P.C. (2011) 'An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies', *Multivariate Behavioral Research*, 46(3), pp. 399–424. Available at: https://doi.org/10.1080/00273171.2011.568786.

Brunner, E., Bathke, A. and Konietschke, F. (2018) *Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs*. Springer Cham (Springer Series in Statistics (SSS)).

Bucher, H.C. *et al.* (1997) 'The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials', *Journal of Clinical Epidemiology*, 50(6), pp. 683–691. Available at: https://doi.org/10.1016/S0895-4356(97)00049-8.

Büning, H. (2002) 'An adaptive distribution-free test for the general two-sample problem', *Computational Statistics*, 17(2), pp. 297–313. Available at: https://doi.org/10.1007/s001800200108.

Casella, G. and Berger, R.L. (2002) *Statistical Inference*. Thomson Learning.

Chen, J. *et al.* (2021) 'The Current Landscape in Biostatistics of Real-World Data and Evidence: Clinical Study Design and Analysis', *Statistics in Biopharmaceutical Research* [Preprint]. Available at: https://doi.org/10.1080/19466315.2021.1883474.

Cheng, D., Ayyagari, R. and Signorovitch, J. (2020) 'The statistical performance of matching-adjusted indirect comparisons: Estimating treatment effects with aggregate external control data', *The Annals of Applied Statistics*, 14(4), pp. 1806–1833. Available at: https://doi.org/10.1214/20-AOAS1359.

Eichler, H.-G. *et al.* (2016) '"Threshold-crossing": A Useful Way to Establish the Counterfactual in Clinical Trials?', *Clinical Pharmacology & Therapeutics*, 100(6), pp. 699–712. Available at: https://doi.org/10.1002/cpt.515.

Eichler, H.-G. *et al.* (2020) 'Are Novel, Nonrandomized Analytic Methods Fit for Decision Making? The Need for Prospective, Controlled, and Transparent Validation', *Clinical Pharmacology and Therapeutics*, 107(4), pp. 773–779. Available at: https://doi.org/10.1002/cpt.1638.

EMA (2018) *Orphan designation: Overview*, *European Medicines Agency*. Available at: https://www.ema.europa.eu/en/human-regulatory/overview/orphan-designation-overview (Accessed: 3 August 2022).

EUnetHTA (2021) *Project Plan D.4.6 - Validity of Clinical Studies - Version 1.0.* Available at: https://www.eunethta.eu/wp-content/uploads/2021/12/EUnetHTA-21-D4.6-Validity-of-clinical-studies-Project-Plan-v1.0.pdf (Accessed: 3 September 2022).

European Parliament (1999) *Regulation (EC) No 141/2000 of the European Parliament and of the Council of 16 December 1999 on orphan medicinal products*, *OJ L*. Available at: http://data.europa.eu/eli/reg/2000/141/oj/eng (Accessed: 25 July 2022).

Ford, I. and Norrie, J. (2016) 'Pragmatic Trials', *New England Journal of Medicine*, 375(5), pp. 454–463. Available at: https://doi.org/10.1056/NEJMra1510059.

Franklin, J.M. and Schneeweiss, S. (2017) 'When and How Can Real World Data Analyses Substitute for Randomized Controlled Trials?', *Clinical Pharmacology and Therapeutics*, 102(6), pp. 924–933. Available at: https://doi.org/10.1002/cpt.857.

G-BA (2008) 'Verfahrensordnung des Gemeinsamen Bundesausschusses in der Fassung vom 18. Dezember 2008 zuletzt geändert durch den Beschluss vom 19. Mai 2022'. Available at: https://www.g-ba.de/richtlinien/42/ (Accessed: 3 August 2022).

G-BA (2022a) *Justification of the Resolution of the Federal Joint Committee (G-BA) on an Amendment of the Pharmaceuticals Directive: Annex XII – Benefit Assessment of Medicinal Products with New Active Ingredients according to Section 35a SGB V Idecabtagen vicleucel (multiple myeloma, at least 3 prior therapies)*. Benefit Assessment. Available at: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/781/#dossier (Accessed: 13 July 2022).

G-BA (2022b) *Nutzenbewertung - Bewertung von Arzneimitteln für seltene Leiden nach § 35a Absatz 1 Satz 11 i.V.m. 5. Kapitel § 12 Nr. 1 Satz 2 VerfO, Wirkstoff: Idecabtagen vicleucel*. Benefit Assessment. Available at: https://www.g-ba.de/bewertungsverfahren/nutzenbewertung/781/#dossier (Accessed: 13 July 2022).

Guyatt, G.H. *et al.* (2008) 'GRADE: an emerging consensus on rating quality of evidence and strength of recommendations', *BMJ*, 336(7650), pp. 924–926. Available at: https://doi.org/10.1136/bmj.39489.470347.AD.

Hernán, M.A. (2014) 'Confounding - Structure in Wiley StatsRef: Statistics Reference Online', in *Wiley StatsRef: Statistics Reference Online*. John Wiley & Sons, Ltd. Available at: https://doi.org/10.1002/9781118445112.stat03729.

Hernán, M.A. and Robins, J.M. (2016) 'Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available', *American Journal of Epidemiology*, 183(8), pp. 758–764. Available at: https://doi.org/10.1093/aje/kwv254.

Hoffmann, F. *et al.* (2021) 'Versorgungsnahe Daten zur Evaluation von Interventionseffekten: Teil 2 des Manuals', *Das Gesundheitswesen*, 83(06), pp. 470–480. Available at: https://doi.org/10.1055/a-1484-7235.

ICH E9 (1998) *E9 Statistical Principles for Clinical Trials*. Guideline. Available at: https://database.ich.org/sites/default/files/E9_Guideline.pdf (Accessed: 12 July 2022).

ICH E9 R1 (2019) *E9(R1) Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles in Clinical Trials*. Guideline. Available at: https://database.ich.org/sites/default/files/E9-R1_Step4_Guideline_2019_1203.pdf (Accessed: 12 July 2022).

IQWiG (2020) *Konzepte zur Generierung versorgungsnaher Daten und deren Auswertung zum Zwecke der Nutzenbewertung von Arzneimitteln nach § 35a SGB V*. Rapid Report A19-43, Nr. 863.

IQWiG (2022a) *Allgemeine Methoden Version 6.1 vom 24.01.2022*. Available at: https://d-nb.info/1251569048/34 (Accessed: 22 August 2022).

IQWiG (2022b) *Versorgungsnahe Daten in Herstellerdossiers: Es läuft noch nicht rund*. Available at: https://www.iqwig.de/presse/pressemitteilungen/pressemitteilungen-detailseite_67103.html (Accessed: 8 August 2022).

Johnson, N.L., Kotz, S. and Balakrishnan, N. (1995) *Continuous Univariate Distributions, Volume 2*. 2nd edn. New York: Wiley-Interscience.

Leverkus, F. and Chuang-Stein, C. (2016) 'Implementation of AMNOG: An industry perspective', *Biometrical Journal*, 58(1), pp. 76–88. Available at: https://doi.org/10.1002/bimj.201300256.

Morris, T.P., White, I.R. and Crowther, M.J. (2019) 'Using simulation studies to evaluate statistical methods', *Statistics in Medicine*, 38(11), pp. 2074–2102. Available at: https://doi.org/10.1002/sim.8086.

Patel, D. *et al.* (2021) 'EC in SAT - Use of External Comparators for Health Technology Assessment Submissions Based on Single-Arm Trials', *Value in Health*, 24(8), pp. 1118–1125. Available at: https://doi.org/10.1016/j.jval.2021.01.015.

Pearl, J. (1995) 'Causal Diagrams for Empirical Research', *Biometrika*, 82(4), pp. 669–688. Available at: https://doi.org/10.2307/2337329.

Phillips, B. (2014) *The crumbling of the pyramid of evidence*, *ADC Online Blog*. Available at: https://blogs.bmj.com/adc/2014/11/03/the-crumbling-of-the-pyramid-of-evidence/ (Accessed: 3 August 2022).

Piantadosi, S. (2005) *Clinical Trials: A Methodologic Perspective: Second Edition*. Wiley-Blackwell. Available at: https://doi.org/10.1002/0471740136.

Rencher and Schaalje (2008) *Linear Models in Statistics, 2nd Edition*. Available at: https://www.wiley.com/en-us/Linear+Models+in+Statistics%2C+2nd+Edition-p-9780471754985 (Accessed: 8 August 2022).

Ribeiro, T.B. *et al.* (2022) 'Single-arm clinical trials that supported FDA Accelerated Approvals have modest effect sizes and at high risk of bias', *Journal of Clinical Epidemiology*, 0(0). Available at: https://doi.org/10.1016/j.jclinepi.2022.01.018.

Rubin, D.B. (1974) 'Estimating causal effects of treatments in randomized and nonrandomized studies.', *Journal of Educational Psychology*, 66(5), pp. 688–701. Available at: https://doi.org/10.1037/h0037350.

Schouten, H.J. (1999) 'Sample size formula with a continuous outcome for unequal group sizes and unequal variances', *Statistics in Medicine*, 18(1), pp. 87–91. Available at: https://doi.org/10.1002/(sici)1097-0258(19990115)18:1<87::aid-sim958>3.0.co;2-k.

Searle, S.R. and Gruber, M.H.J. (2016) *Linear Models*. John Wiley & Sons.

Senn, S. (1997) *Statistical Issues in Drug Development*. Chichester: Wiley (Statistics in practice).

Signorovitch, J.E. *et al.* (2010) 'Comparative Effectiveness Without Head-to-Head Trials', *PharmacoEconomics*, 28(10), pp. 935–945. Available at: https://doi.org/10.2165/11538370-000000000-00000.

Sterne, J.A. *et al.* (2016) 'ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions', *BMJ*, p. i4919. Available at: https://doi.org/10.1136/bmj.i4919.

Tennant, P.W.G. *et al.* (2021) 'Use of directed acyclic graphs (DAGs) to identify confounders in applied health research: review and recommendations', *International Journal of Epidemiology*, 50(2), pp. 620–632. Available at: https://doi.org/10.1093/ije/dyaa213.

Appendix

In the following the derivation of reweighted treatment mean and ESS based on individual weights is given. First note that:

$$\omega_{i.mild} = \frac{\pi_{hist}}{\pi_{treat}} * \frac{1}{n_{treat}} \quad , mildly\ diseased\ patients$$

$$\omega_{i,sev} = \frac{1-\pi_{hist}}{1-\pi_{treat}} * \frac{1}{n_{treat}} \quad , severely\ diseased\ patients$$

$$n_{treat,mild} = n_{treat} * \pi_{treat}$$

$$n_{treat,sev} = n_{treat} * (1-\pi_{treat})$$

Then it can be calculated that individual weights chosen as above corresponds to the aggregate weights used in this thesis.

$$\begin{aligned}
\bar{y}_{treat(hist)} &= \sum_{i=1}^{n_{treat}} \omega_i * y_{i,treat} \\
&= \sum_{i=1}^{n_{treat,mild}} \omega_{i.mild} * y_{i,treat,mild} + \sum_{i=1}^{n_{treat,sev}} \omega_{i,sev} * y_{i,treat,sev} \\
&= \sum_{i=1}^{n_{treat,mild}} \frac{\pi_{hist}}{\pi_{treat}} \frac{1}{n_{treat}} * y_{i,tr,mild} + \sum_{i=1}^{n_{treat,sev}} \frac{1-\pi_{hist}}{1-\pi_{treat}} \frac{1}{n_{treat}} * y_{i,tr,sev} \\
&= \pi_{hist} * \frac{1}{n_{tr,mild}} \sum_{i=1}^{n_{treat,mild}} y_{i,tr,mild} + 1-\pi_{hist} * \frac{1}{n_{tr,mild}} \sum_{i=1}^{n_{treat,sev}} y_{i,tr,sev} \\
&= \pi_{hist} * \bar{y}_{treat|mild} + (1-\pi_{hist}) * \bar{y}_{treat|severe}
\end{aligned}$$

Additionally, the sample size formula can be derived:

$$ESS = \frac{(\sum_{i=1}^{n_{treat}} \omega_i)^2}{\sum_{i=1}^{n_{treat}} {\omega_i}^2} = \frac{\left(n_{treat,mild} * \omega_{i.mild} + n_{treat,sev} * \omega_{i.sev}\right)^2}{n_{treat,mild}(\omega_{i.mild})^2 + n_{treat,sev}(\omega_{i.sev})^2}$$

$$= \frac{\left(n_{treat,mild}\frac{\pi_{hist}}{\pi_{treat}} * \frac{1}{n_{treat}} + n_{treat,sev}\frac{1-\pi_{hist}}{1-\pi_{treat}} * \frac{1}{n_{treat}}\right)^2}{n_{treat,mild}\left(\frac{\pi_{hist}}{\pi_{treat}} * \frac{1}{n_{treat}}\right)^2 + n_{treat,sev}\left(\frac{1-\pi_{hist}}{1-\pi_{treat}} * \frac{1}{n_{treat}}\right)^2}$$

$$= \frac{\left(\pi_{treat} * n_{treat} * \frac{\pi_{hist}}{\pi_{treat}} * \frac{1}{n_{treat}} + (1-\pi_{treat}) * n_{treat}\frac{1-\pi_{hist}}{1-\pi_{treat}} * \frac{1}{n_{treat}}\right)^2}{\pi_{treat} * n_{treat}\left(\frac{\pi_{hist}}{\pi_{treat}} * \frac{1}{n_{treat}}\right)^2 + (1-\pi_{treat}) * n_{treat}\left(\frac{1-\pi_{hist}}{1-\pi_{treat}} * \frac{1}{n_{treat}}\right)^2}$$

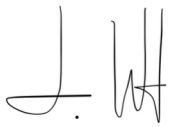
$$= \frac{(\pi_{hist} + 1 - \pi_{hist})^2}{\frac{{\pi_{hist}}^2}{\pi_{treat} * n_{treat}} + \frac{(1-\pi_{hist})^2}{(1-\pi_{treat}) * n_{treat}}} = \frac{n_{treat}}{\frac{{\pi_{hist}}^2}{\pi_{treat}} + \frac{(1-\pi_{hist})^2}{(1-\pi_{treat})}} = \frac{1}{\lambda} * n_{treat}$$

Affidavit

I hereby confirm that my thesis entitled “Threshold-Crossing for Single Arm Trials with External Control in Form of Aggregate Data” is the result of my own work. All sources and / or materials applied are listed and specified in the thesis.

Furthermore, I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form.

Göttingen, 04. September 2022
Place, Date

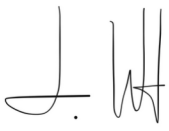

Signature

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, die Masterarbeit „Threshold-Crossing für Einarmige Studien mit Externen Kontrollen in Form Aggregierter Daten“ eigenständig angefertigt und keine anderen als die von mir angegebenen Quellen und Hilfsmittel verwendet zu haben.

Ich erkläre außerdem, dass die Masterarbeit weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.

Göttingen, 04. September 2022
Ort, Datum


Unterschrift