Ludwig-Maximilians-Universität München
Institute for Medical Information Processing, Biometry, and Epidemiology (IBE)
Chair of Medical Biometry with Focus on Clinical Epidemiology

Bachelor Thesis

Phase II Studies in Oncology: Current Practices of Choosing Group Configuration and Statistical Design

by Katja Gutmair
supervised by Prof. Dr. Eva Hoster
July 12, 2021

Abstract

Phase II studies take a central role in the process of testing drug candidates in oncology because they evaluate the drug candidates for efficacy (phase IIa) and select promising drug candidate for phase III studies. (phase IIb) Phase II studies are usually open-label, single-arm studies with a binary endpoint, a historical control, and one or two stages. The recent development of targeted and immunotherapeutic drugs instead of cytotoxic drugs changed the demands on a phase II study. For addressing these demands, a multitude of designs for a phase II study has been developed.

There are many variations in the design of a phase II study. The choice of the primary endpoints depends on the property of the drug candidate. Statistical inference can be done either with the frequentist inference (hypothesis-testing) or with the Bayesian inference, which is based on a priori and a posteriori probability. A single-arm study uses a historical control as comparison for the efficacy of a drug. Especially for targeted and immunotherapeutic drugs, suitable historical controls are not available. In this case, a two-or multi-arm study, with possible randomization between these arms, may be the appropriate choice, in which one drug is compared to the standard therapy or in which only drug candidates are compared to each other without a control arm to select the most efficacious drug candidate. Other variations in the conduction of a phase II study include interim evaluations, and seamless phase I/II and phase II/III studies.

Many designs have been developed, which incorporate these the most common designs used are Fleming´s single-stage design, Fleming´s two- or multistage design, Gehan´s single-stage design, and Simon´s two-stage Optimum and Minimax Design.

A literature review in selected peer-reviewed journals as well as in EudraCT for phase I/II, phase II, and phase II/III studies in oncology limited to the years 2019/2020 was done. Information about group configuration and statistical designs were extracted to evaluate differences in theoretically proposed and practically used designs. 60 recently published studies found in selected peer-reviewed studies and 43 recently approved studies found on EudraCT could be extracted. There were only a few phase I/II and phase II/III studies. Most of all reviewed studies were designed as single-arm (recently published studies: 50%, recently approved studies: 58%) or as two-arm studies (recently published studies: 42%, recently approved studies: 31%). 97% of recently published and all recently approved two- or multi-arm studies randomized between the arms. 67% of recently published studies did not mention a specific statistical design Most common designs used were Simon´s two-stage Optimal (8%) and Minimax Design (7%) and O´Brien Fleming Design (7%). Hypotheses were well described, but information provided about the statistical test for sample size calculation was insufficient

There are many theoretical proposed designs for different drug candidates, endpoints, and aims. However, in practice, only a few of them are used. A reason for this might be, that many theoretical proposed designs are complex, complicated to understand, and therefore not widely used. Another reason might be lack of official guidelines for conduction phase II studies in oncology. Therefore, better education and guidelines could lead to an improvement in the design and statistical analysis of phase II studies.

Content

1.Introduction

Before a new drug in oncology can be used as a therapeutic intervention for patients, its safety and efficacy must be ensured in several studies. Usually, a drug candidate must pass three study phases before market approval. phase I studies are carried out in a small study population and determine the maximum tolerated dose limited by toxicity. The aim of phase II studies is to determine the efficacy of a drug candidate, and to gain more information about the safety and toxicity (phase IIa “proof of concept”), and to identify and select promising agents for phase III studies. (phase IIb). Phase III is the last phase before market approval. It aims to determine if the candidate drug is better than the standard therapy. Phase III studies are usually large, multicenter, randomized studies and therefore expensive and time-consuming. (Winter and Pugh 2019)

For a long time, there were only a few chemotherapeutic drugs with a high level of cytotoxicity available for clinical studies. This changed over the past decade. Due to new advances in molecular biology, the available number of new possible antitumor drugs has increased. These targeted agents do mostly not aim at cytotoxicity but on the modification of mutated cellular pathways, which are often the reasons for abnormal cellular growth. Along with these new drug candidates, some problems and challenges in clinical testing arise. One problem is that there are now numerous drug candidates available for clinical testing, but patients participating in clinical studies are limited. Furthermore, the number of patients receiving a potentially inactive drug should be kept as low as possible for ethical reasons. Other problems are limited financial resources and the need for different endpoints because targeted drugs aim at tumor size stabilization rather than tumor size reduction. (Farley and Rose 2010; Cannistra 2009)

In the drug development process phase II studies are an important key element because this phase tests a drug candidate in terms of efficacy and selects drug candidates for phase III testing if its efficacy seems to be high enough compared with the standard treatment or an historical control. (Lee and Feng 2005) Nowadays only 58% of all phase III studies are successful and only 15% of all drugs entering a phase II study will gain market approval. (Thomas et al. 2016) As a consequence, the demands on a phase II study are to identify effective drugs and reject ineffective drugs correctly and as quickly as possible with a minimum number of patients to reduce the number of drugs failing in a Phase III study and therefore reduce costs and number of patients receiving an inactive drug. (Lee and Feng 2005)

The standard design of a phase II study in oncology is an open-label single-arm study with a binary endpoint, one or two stages, and a historical control. The sample size usually includes 35 – 50 persons and the duration is usually 18 – 24 months with longer follow-ups. (Schlesselman and Reis 2006) In the last two decades, numerous other

designs have been developed, that address the demands of a phase II study. Such designs differ in the usage of endpoints, the number of stages and treatment arms, randomization, adaptive designs, and different statistical inference frameworks. Recent approaches suggest combining phases resulting in phase I/II and phase II/III designs. (Hess 2007; Ang et al. 2010)

This thesis aims on giving an overview over group configuration and theoretical statistical designs for phase II studies in oncology. Furthermore, this work deals with the question, which group configurations and statistical approaches are used in practice. Differences between theoretical proposals and practical usages are discussed.

The outline of this thesis is as follows: In Section 2 the major characteristics of a phase II study including possible endpoints, number of stages and arms and inference based on the frequentist and Bayesian setting are explained. Furthermore, principles of the adaptive design and combined phase I/II and phase II/III are explained. The most popular statistical designs used in a phase II study are presented in Section 3. To evaluate, which of these described statistical designs are used in practice, a literature review dealing with this question is conducted in Section 4. In Section 5 the results of Section 4 are discussed and compared with the theoretical statistical designs. In section 6, a conclusion is drawn.

2. Design and statistical analysis of phase II studies

There are a multitude of different designs and statistical analysis for a phase II study in oncology. Common variations of a phase II design include the choice between different endpoints, statistical inference, number of drugs and treatment arms and interim evaluation for the possibility of an early stopping of the study. The choice of these characteristics depends on the purpose of the study, and available financial and patient resources. In recent years, there is an increased interest in randomization and adaptive designs hoping that these practices can lead to more reliable results. A newer approach is to break up the relative inflexible sequence of phase I, phase II, phase II in favor of seamless phase I/II and phase II/III designs. Additionally, there is an ongoing debate what kind of reference should be used for evaluating the superiority of the drug candidate. (Ang et al. 2010; Sargent and Taylor 2009)
The following section provides an overview over the common design variations and its recommended application.

2.1 Endpoints

Choosing the suitable endpoint is essential for reducing time, costs, and incorrect decisions e.g., sending a drug candidate to phase III although it has not enough tumor activity or vice versa. Therefore, an endpoint should be sensitive, reproducible, specific, well defined, of clinical relevance, objective and not expensive to measure. In phase II clinical studies in oncology, there is usually one primary outcome, that evaluates the efficacy of one or several drug candidates, and several secondary outcomes for additional outcomes, e.g., toxicity, and safety. (Kilickap et al. 2018; Ellimoottil et al. 2015)
Objective response rate (ORR) has been historically the most popular primary endpoint. This has been changed in recent years in favor of increasing usage of overall survival (OS) and progression-free survival (PFS). This change is due to the fact of the development and testing of targeted and immunotherapeutic drugs or a combination of both instead of just using chemotherapeutic drugs. These new treatment approaches lengthen patient´s survival and aim at tumor size stabilization rather than tumor size reduction, so that ORR as primary endpoint may not be practical and appropriate anymore. (Kilickap et al. 2018; Wu et al. 2011) The most common endpoints are described in this section. ***Table 1*** provides an overview, which endpoint is recommended for which type of a phase II study.

Table 1: Recommendations which endpoint is appropriate for which kind of phase II study (Wu et al. 2011, Dhani et al. 2009)

Endpoint	Recommended for
Objective response rate	Studies with short running time, small study population and cytostatic agents with expected relatively high and fast activity in tumor shrinkage Multicenter studies
Progression-free survival	Studies with limited time and patient number (for greater study population also appropriate) and drugs not only aiming on tumor size reduction but tumor size stabilization Appropriate surrogate marker for OS Appropriate for gaining information about drug activity Appropriate for studies with a possible cross over effect or subsequent therapies May not appropriate for drugs with long time period till efficacy
Disease-free survival	For studies evaluation the benefit of adjuvant therapies Tumor-recovered patients with no relapse on study start
Time to progression	Not widely used in clinical studies because, death as event is excluded
Overall survival	Gold standard in Phase II studies Requires large study population and a cohort as control High financial costs Long duration till outcome is available Therefor not recommended in a phase II study, that aims on a result in a short period of time
biomarkers	High potential in gaining additional information about mechanism of resistance and as predictive outcome Only a few validated biomarkers available for a specific cancer type Therefore, recommended as secondary endpoint
Pathological complete response	Neoadjuvant studies Studies with demand on accelerated approval
Quality of Life	secondary endpoint because of subjectivity primary endpoint in palliative therapeutic agents

Objective Response Rate

The objective response rate measures the number of patients achieving a tumor size reduction or tumor disappearance, measured as partial response (PR) or complete response (CR), in proportion to all patients participating in the study for a minimum time period. CR is the lack of a detectable tumor and PR is the reduction of tumor size of a predefined amount. (FDA 2018) ORR is a standardized outcome defined first by the World Health Organization (WHO) and then by Response Evaluation Criteria in Solid Tumors (RECIST). In clinical practice, the definition of ORR is often used by RECIST criteria. The detailed response criteria in solid tumors are listed in ***Table 2***. (Wu et al. 2011) In most studies, a drug candidate is declared as efficacious if the response rate improves about 15% to 20% compared with the standard therapy (Digmam et al. 2006). ORR is an appropriate measurement of outcome for studies with a short running time, small sample size, and cytotoxic agents because tumor shrinkage induced by cytostatic drugs can be observed after a short period of time. ORR may not be suitable for phase II studies, that examine the efficacy of a targeted or immunotherapeutic drug candidate, because these drugs aim at tumor size stabilization or deceleration of tumor growth rather than tumor shrinkage. A concern with using this endpoint is, that high ORR is not always correlated with longer survival. Another concern of ORR as a clinical endpoint is that tumor shrinkage may occur later due to the mechanism of action of the drug and the study may not be long enough to capture this outcome. Furthermore, patients receive these new drug candidates over a longer period compared to chemotherapeutic drugs. The use of ORR can lead to a loss of information because the continuous variable “tumor shrinkage” is categorized. To summarize, provided, that ORR is an appropriate surrogate endpoint for survival or PFS, it is an appropriate endpoint for cytostatic drug candidates, that lead to a fast tumor size reduction, but for the evaluation of targeted and immunotherapeutic agents, time-related endpoints may be a better choice. (Wu et al. 2011, Dhani et al. 2009)

Table 2: Response criteria in solid tumors according to RECIST guidelines (Eisenhauer et al. 2009)

endpoint	RECIST criteria
CR	Disappearance of all target lesions Reduction of pathological lymph nodes in short axis to < 10 mm
PR	≥ 30 % reduction in the sum of diameter of target lesions Reference: baseline sum diameter
PD	≥ 20 % increase in the sum of diameters of target lesions and an absolute increase of ≥ 5 mm Reference: smallest sum on study
SD	Shrinkage or decrease neither in PR not in PD met Reference: smalless rum of diameters

Time related endpoints

There is an increasing interest in using time-related endpoints in phase II studies. PFS describes the time from initiation of a therapy to disease worsening, that includes tumor growth as well as the development of new lesions. DFS defines the time from complete tumor disappearance to tumor relapse. The definition of time to progression (TTP) is the same as the definition of PFS but death is not included as an event in TTP. Usually, TTP is not widely used in clinical studies.

A concern of the use of time-related endpoint is that not all patients will have a disease worsening during the study. This issue can be cleared by censoring these patients using Kaplan Meier estimation. Log-rank tests are used for comparing two or more time-to-event curves in a study with two or more arms and drugs. Time-related endpoints are more detailed and provide more information than ORR by measuring the duration till disease worsening instead of categorizing tumor shrinkage into two categories. Disadvantages of time-related endpoints are, that they can be biased due to frequent evaluations.

Compared to ORR, PFS is easy to measure and an appropriate endpoint for gaining information about drug activity. Because of the fact, that disease progression occurs earlier than death, PFS is not as time-consuming as OS and may be therefore a suitable surrogate marker for OS. Compared to OS, PFS is not biased by subsequent therapies. An advantage of PFS compared to ORR is, that PFS includes the measurement of stable disease. Compared to OS, this endpoint requires only a small number of patients. A disadvantage is, that a long PFS does not always lead extended survival. It also shows

some weakness in studies using immunotherapies, which may be related to the fact, that immunotherapies show evidence just after 10 – 12 weeks.
While PFS is the appropriate endpoint for evaluating the efficacy of an anti-tumor drug in ill patients, DFS is used in studies that evaluate the benefit of adjuvant therapies of tumor-recovered patients with no relapse on study start. (Wu et al. 2011; Kilickap et al. 2018)

Overall Survival

OS measures the time from initial treatment to death from any cause. An advantage of OS is, that this outcome is simple to measure and clinically relevant because survival is the most reliable cancer endpoint. (FDA 2018) OS is not prone to evaluation or researcher bias. Because of this, OS is seen as the gold standard of clinical endpoints. But this endpoint is time-consuming and requires long-term follow-up, which goes along with a need for higher patient numbers leading to higher costs. (Wu et al. 2011, Kilickap et al. 2018) Patients, who did not die before the time of evaluation, get censored (Delgado 2021). Consequently, it is not the first choice in phase II studies in most cases. OS as outcome may be inappropriate, if the patients receiving additional drugs off-study because this could lead to carry-over-effect resulting in a biased OS. In this case, time-related endpoints like PFS, DFS and TTP are the appropriate choice, because they are not influenced by carry-over effects. OS as outcome might be not the preferred choice by antineoplastic therapies and slowly progressive disease, because its life-prolonging properties could mask the efficacy resulting from the testing drug. So, OS seems to be a suitable endpoint for studies with a low survival time and no subsequent therapies option. (Wu et al. 2011, Kilickap et al. 2018)

Biomarkers as endpoints

Surrogate endpoints, e.g., biomarkers can be an appropriate substitute for the above-mentioned clinical endpoints, especially, if the above-mentioned endpoints are non-reliable outcomes for the anti-tumor effect. Before a biomarker can be used as a surrogate endpoint, its evidence must be proofed in studies. Strengths of surrogate endpoints are, that they are less expensive, and they can be measured earlier and more frequently than clinical endpoints. Biomarkers as endpoint could be contributing to an increasing understanding of resistance mechanisms. Surrogate endpoints are tumor-specific, therefore a surrogate endpoint may be a reliable substitute for a clinical endpoint in a certain tumor type, but not in another tumor type. (Wu et al. 2011; Dhani et al. 2009)

Other endpoints

Pathological response rate

Pathological complete response (pCR) is not a common endpoint. Its applications can be found in neoadjuvant studies and accelerated drug approvals. It directly measures the decrease in tumor diameter and can therefore show drug efficacy in a short period of time. (Kilickap et al. 2018)

Quality of Life

Quality of Life (QoL) is a subjective outcome relying on the patient´s benefit. Because of its subjectivity, it is used more as a secondary endpoint rather than a primary endpoint. QoL as primary endpoint is used in palliative therapeutic agents and as an indirect measure for toxicity. (Kilickap et al. 2018)

2.2 Inference of Phase II Studies

In phase II studies, the most common statistical analysis is the frequentist framework. This framework, also called hypotheses testing, tests, if the observed outcome agrees more with the Null-Hypotheses or the alternative Hypotheses. Another inference framework is the Bayesian framework, in which decisions are made due to prior and posterior probabilities. (Perrone et al. 2003) This section provides an overview of the statistical design of these frameworks in phase II studies.

Frequentist inference

The frequentist inference is widely used for decision-making in phase II studies. This inference sees the true response rate as an unknown, but fixed value. Let π the true response rate of the drug candidate. The probability $\mathrm{P}(\mathrm{data}\,|\pi)$ is calculated by collecting these data in a study. (Lee and Chu 2012) This inference is based on two hypotheses, the Null-Hypotheses (H_0) and the alternative Hypotheses (H_1). Let π_0 be the maximum response rate for which the drug candidate is declared as not sufficient efficacious and π_1 the minimum response rate, for with the drug candidate is declared as sufficient efficacious. The value of π_0 is usually the response rate of the standard therapy. The value of π_1 is set to $\pi_0 + \Theta$, where Θ describes the smallest acceptable improvement of the response rate of the drug candidate, also known as clinically relevant improvement,

compared to the standard therapy. (Winter and Pugh 2019; Rubinstein et al. 2009) The value of π_1 must be specified for power calculation and therefore, for sample size calculation. Studies, that investigate, if the drug candidate is superior to standard therapy, π_0 is usually set as followed: $\pi_0 = \text{response rate for standard therapy} - \Theta/2$. The value of π_0 is smaller than the response rate of the standard therapy to ensure, that the drug candidate is moving on to stage 2, even if the drug candidate´s response rate is only slightly better than the available therapy. Especially drug candidates with a new mechanism of action may be efficacious, although its response rate is below the response rate of the standard therapy. (Schlesselman et al. 2006)
In phase II studies, the drug candidate is usually declared as superior, if an improvement of $\Theta = 15\%$ or $\Theta = 20\%$ compared to the standard treatment is observed. Finally, let π the true response rate of the drug candidate. The response rate receiving in a study serves as an estimator for π. Given these parameters, the hypotheses are usually formulated as follows:

$$H_0: \pi \leq \pi_0$$
$$H_1: \pi > \pi_0.$$

(Schlesselman et al. 2006; Dignam et al. 2006) Besides these parameters, the values of the false-positive error rate denoted as type I error level and false-negative error rate denoted as type II error level must be defined. Most statistical designs proposed for single-arm one-stage, single-arm multi-stage, or multi-arm multistage use this inference, e.g., Fleming´s single-stage designs, Simon two-stage Optimum or Minimax Design. (Perrone et al. 2003) To ensure adequate power (1 – type II error level), that is the probability of correctly rejecting the Null-Hypothesis, adequate sample size is calculated based on the smallest acceptable improvement of the response rate of the drug candidate, Θ, false-positive error rate, and false-negative error rate. In phase II designs, the false negative rate is usually set at 10 % to 20%. The false-positive is usually set at 5% to 10%. Based on the p-value or confidence interval, the Null-Hypotheses can be accepted or rejected. (Winter and Pugh 2019; Rubinstein et al. 2009)

Bayesian Inference

The Bayesian Inference is calculated as follows: $P(\pi \mid \text{data})$ (Lee and Chuh 2012). In contrast to the frequentist inference, the unknown true response rate is seen here as random following a certain probability distribution, denoted as a priori distribution. Depending on its uncertainty and available information, this distribution has a certain shape. If little knowledge about the parameter is given, its prior distribution is flat, also

called “weakly informative” prior. Much knowledge about the parameter Θ resulting in a “high informative” distribution. After the conduction of the study and therefore after receiving a value for the response rate of the drug candidate, this a priori distribution can be updated due to the newly gained information about the response rate of the patients. This updated a priori distribution is then called a posteriori distribution. (Perrone et al. 2003, Ang et al. 2010)

The a posteriori distribution is calculated as follows:

$$P(\pi \mid data) \approx P(\pi)\, P(data \mid \pi)$$

with a priori distribution $P(\pi)$ and the likelihood $P(data \mid \pi)$. (Lee and Chu 2012).

With this given distribution, the probability of observing a certain response rate given a certain number of patients can be calculated. For decision-making, this calculated probability is compared with the observed response rate of the study.

To sum up, in contrast to the frequentist inference, decisions are not made based on p-values or confidence intervals, but on a posteriori probability, prediction- and credibility intervals

The two-stage single-threshold design (STD) or the dual-threshold design (DTD) proposed by Tan and Machin use the Bayesian inference for decision making. Sample size calculation is done by assuming, that the true response rate π is greater than a target value. In a STD, only an upper threshold for the target value is defined, in an STD, a lower and an upper threshold is defined, respectively. (Tan and Machin 2002).

Another design based on the Bayesian inference with early stopping rules was proposed by Thall and Simon. Let π_s the true response rate of the standard therapy and π_e the true response rate for the drug candidate. A “moderate informative” prior distribution is assigned to π_s and a “weakly informative” prior distribution to π_e. The posteriority probability, that π_s is greater than π_e by a certain difference Θ, is updated after accumulating information. Patients’ enrollment is ongoing, till inferiority or superiority of the drug candidate is shown by a low or high posteriority probability. In case of reaching the calculated maximal sample size, denoted by X_{max}, without proof of superiority or inferiority, the studies are declared as inconclusive. (Thall and Simon 1994)

Heitjan proposed a Bayesian design, in which the superiority of the drug candidate is seen with skeptics, and the inferiority of the drug candidate is seen enthusiastic. Different prior probabilities are used to express this scepsis or enthusiasm. After the observed outcome, its posteriori probabilities are calculated, resulting in a “persuade-the pessimist” probability based on the skeptic a priori and a “persuade-the-optimist probability” based on the enthusiastic a priory. If none of these probabilities are high enough, the study is declared inconsistent. (Heitjan 1997)

Advantages and disadvantages

The frequentist framework is easy to understand and rigor, but also relative inflexible. This may lead to a large sample size to reach a certain level of power. Another problem of this framework is that deviation of the original study design may result in higher error rates. (Ang et al. 2010)

Designs using Bayesian inference are appropriate for adaptive designs, or for designs with early stopping options, or for adding or dropping a treatment arm. They enable a more frequent monitoring and interim decision-making because of its nature to update the a priori probability distribution with accumulating data. In contrast to frequentist interim analysis and group sequential design, the number and timing of interim analysis do not have any impact on the outcome based on Bayesian inference. Depending on the, often subjective information, that is used to calculate the prior distribution, this prior distribution has different characteristics. This may lead to a bias in the inference because based on different a priori distributions, results from the same data might differ. To account for this issue, a priori distributions should be specified in the study protocol. (Lee and Chu 2012

The choice between a frequentist framework and a Bayesian framework depends on the study design as well as tumor type, available financial and patients´ resources, number of treatment arms, and prior knowledge about the drug candidate and tumor type. (Ang et al. 2010)

2.3 Dual endpoints

Most studies in phase II use just one primary endpoint for decision making, whether the drug candidate seems promising.

Dent et al. proposed a dual-endpoint design with both ORR and early progression rate as primary endpoint. A drug candidate is declared ineffective if the response rate of the drug candidate is below a predefined level or the early progression is over a predefined level. (Dent et al. 2001)

Sill et al proposed a two-stage dual-endpoint design similar to Dent et al. If the drug candidate shows a sufficient response rate or increased PFS, the drug candidate is declared as efficacious. The study is terminated after the first stage if neither an increased response rate nor an increased PFS is seen. (Sill et al. 2012) The sample size required for this dual-endpoint design is just slightly higher than for a single-endpoint design using the same statistical properties (Rubinstein 2014).

Some designs, like the design of Bryant and Day, analyze both efficacy and toxicity simultaneously instead of separately, as usual, by evaluating both efficacy and toxicity as primary endpoint. In this sense, a drug candidate is considered promising, if the response rate is greater and the toxicity rate at least equal to the standard treatment. Let p_r be the true response rate of the drug candidate, p_t the true rate of DLT, and p_{r0} the response rate of efficacy and p_{t0} the DLT of the standard treatment, then the Hypotheses are formulated are follows:

$$\text{Ho: } p_r \leq p_{r0} \text{ or } p_t \geq p_{t0}$$
$$\text{H1: } p_r > p_{r0} \text{ and } p_t < p_{t0}$$

If $p_r > p_{r0}$ or $p_t < p_{t0}$, the drug candidate seems to be promising, otherwise, the drug candidate is rejected. Additionally, the efficacy-toxicity-association is modeled by odds ratios for toxicity among responders to non-responders, denoted as Θ. (Bryant and Day 1995)

Conaway and Petroni proposed a similar design to Bryant and Day, with the difference, that this design accepts greater toxicity in higher response rates and vice versa. (Conaway and Petroni 1995)

Sun et al propose a randomized statistical design, that includes two endpoints: RR and early disease progression with high values of α- and β-errors, which are characteristic for randomized studies. This design includes one interim analysis of RR. (Sun et al. 2009)

2.4 Number of Arms and Drugs

A phase II study can be designed as a single-arm study using a historical control as reference for the efficacy of the drug candidate, or as multiple-arm study in which multiple drug candidates, combinations of drug candidates, or different doses of the same drug candidate are tested. In a multiple-arm study, there is either one arm with a standard treatment as reference for efficacy and experimental arm(s) with one or different drug candidates or there are only experimental arms with different drug candidates and the drug candidates are tested for superiority. (Perrone et al. 2003; Brown et al. 2014) Statistical designs including a single arm are proposed by Gehan (Gehan 1961), Fleming (Fleming 1982), and Simon (Simon 1989). For more details about these designs, see **section 3.** An advantage of a single-arm study using a historical control compared to a multi-arm study is, that only a small number of patients is needed. However, a concern is that the outcome of the historical control may not be compared to the outcome of the study in some circumstances because of several reasons: The number and composition

of the patient population may differ between historical and present studies due to different inclusion criteria, prognostic factors, and inter-institutional variability. New developments in science, e.g., differences in radiological and surgical techniques or the development of targeted and immunotherapeutic therapies aiming at different mechanisms. Another disadvantage of historical controls is, that the primary outcome changed over the past years. Today, the use of time-related endpoints is increasing, but historical controls are frequently based on response rates. Furthermore, historical controls including biomarkers as a primary endpoint are lacking. The use of historical controls is appropriate for studies with only a small patient population available, e.g., studies for rare cancer types, and for studies using response rates as primary endpoints. For studies with time-related endpoints, biomarker as clinical endpoints and enough patients available, a two-arm study containing a control arm may be more appropriate. (Rubenstein 2019, Ang et al. 2010)

2.5 Number of Stages

Phase II studies can be conducted as a single-stage design, two-stage design, or multi-stage design. In a single-stage design, a fixed sample of patients is treated and after a set period, statistical analysis is done, and conclusions are drawn of its results. One- and two-stage designs are commonly used in phase II studies in oncology. (Perrone et al. 2013) Fleming proposed a single-arm one-stage design, which is still used today, **see section 3.1** (Fleming 1982). A disadvantage of a single-stage design is, that the study cannot be stopped earlier, if the drug seems to have sufficient anti-tumor activity or if the drug does not show any anti-tumor activity, which is an ethical dilemma. (Perrone et al. 2013)

Two- and multistage designs solve this problem by allowing interim analysis. In the first stage, a set sample of patients is treated with a drug candidate. After a set time, an interim evaluation is conducted. If the response rate of the drug candidate is above a set limit, the study moves to the second stage with an enrolment of additional patients, see ***Figure 1***. Depending on the number of interim analysis, the study is defined as a two- or multistage design. (Perrone et al. 2013) Gehan first proposed 1961 a two-stage design, that determines the required number of patients for maintaining a predefined power. This design allows early stopping after stage one if the drug candidate seems to be ineffective, otherwise, this study moves on to the second stage, **see section 3.3**. (Gehan 1961) Other popular designs include Fleming´s multistage design (**section 3.2**), which allows early stopping for both drug activity and drug inactivity given a predefined number of responders (Fleming 1982) and Simon´s Optimal or Minmax design (**section 3.4**), which

is an extension of Flemings design by minimizing the sample size under the Null-Hypothesis. (Simon 1989) Ensign et al. proposed a three-stage design, which is a combination of Gehan´s first stage and Simon´s Optimum two-stage design, that allows early stopping if no success is observed in patients of the first stage. Otherwise, the study is continued to the second stage. (Ensign et al 1994) Some two-stage and multistage designs screen not only for efficacy but for toxicity of the drug candidate, as example mentioned the design of Bryant and Day (Bryant and Day 1995), see **section 2.3** or a design by Thall and Simon using Bayesian inference (Thall and Simon 1995) Case and Morgan proposed a design that uses as endpoint survival probabilities (Case and Morgan 2003).

Statistical implementation of two- and multistage design

Let s be the maximum response rate, for with the drug candidate is declared as inefficacious and a is the minimal response rate, for with the drug candidate is declared as efficacious. Let n_1 be the number of patients enrolled in the first stage and n_2 the number of patients enrolled in the second stage. Let r_1 the response rate, measured by clinical outcomes, see **section 2.1** in the first stage and r_2 the response rate observed in the second stage. Depending on the value of r_1, the study will be stopped or continued to the second stage by enrolling additional n_2 patients with a total sample size of $N = n_1 + n_2$ and $R = r_1 + r_2$ the cumulative number of successes observed after the second stage.

Decisions for stopping or continuation after the second stage are based on the value of r_1:

- If $r_1 \leq s$: the study is stopped due to insufficient efficacy of the drug given a certain false-negative error β_1.
- If $r_1 \geq a$: the study is stopped due to sufficient efficacy given a certain false-positive error $\alpha 1$.
- If $s < r_1 < a$: the study will continue to the next stage and enrollment of patients continue. (Kramar et al. 1996)

The scheme for decision making in a multi-stage study is equal to the two-stage, with the difference, that more than one interim evaluation is conducted. The precise notation for a decision about stopping the study or moving to the next stage is described in **section 3.2**.

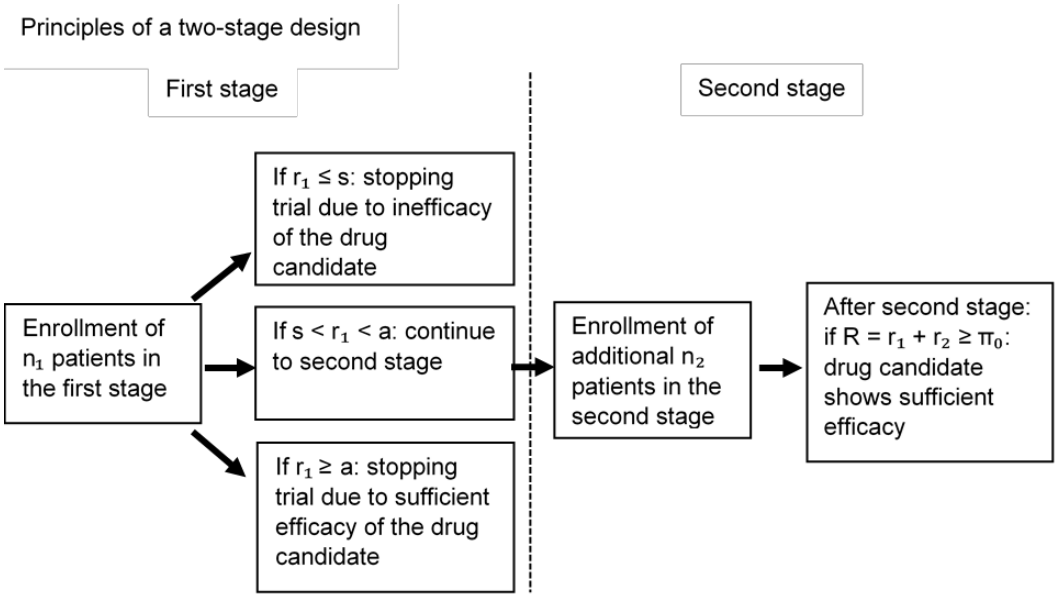

Figure 1: Two stage design. Let r_1 be the number of treatment success observed in stage 1 and r_2 the number of treatment success in stage 2. n_1 is the number of patients enrolled in stage 1 and n_2 the additional number of patients enrolled in stage 2. Let s be the maximum response rate, for with the drug candidate is declared as inefficacious and a is the minimal response rate, for with the drug candidate is declared as efficacious for decision making in the interim analysis. π_o is the response rate under the standard treatment. (Kramar et al. 1996)

Advantages and disadvantages

Multi-stage designs are relatively easy to understand and implement. They reduce the sample size compared to a single-stage design with the same statistical properties. An advantage of the multi-stage design is the higher opportunity to detect an inefficient drug earlier due to a higher number of interim evaluations compared to a two-stage design. This minimizes the likelihood and time, in which patients are treated with a potentially ineffective drug. A difficulty of two- or multistage design is, that the interim evaluation requires an interruption of the study and patients` accrual. This may result in organizational difficulties, especially in a multicenter setting. The usage of adaptive design can avoid this problem of stopping patients´ accrual., see **section 2.7** Designing a multi-stage design is more complex and more difficult to conduct in practice, requires more frequent monitoring and longer duration compared to a single-stage design

Single-stage designs are appropriate in a setting, in which the overall aim is to estimate the efficacy of the drug candidate precisely or in situations with rapid patients' enrollment. Two- or multistage designs are appropriate if the value of the endpoints is available after only a long period and for drugs with a high likelihood of being inactive.

(Perrone et al. 2013; Schlesselman et al. 2006)

2.6 Randomization

There is an increasing demand and use of randomized two- or multiple arms phase II studies because of several factors: The increasing number of available potential anti-tumor drugs makes it impossible to screen every single drug candidate in a single-arm study because of limited patients and financial resources. As mentioned before, another problem is, that historical control may not be reliable or available because of the use of novel endpoints e.g., PFS, biomarkers, or novel drugs like targeted or immunotherapeutic drugs. PFS, which has an increased use in practice, is easily influenced by non-therapeutic-effected factors, that occur in non-homogeneous groups. Instead of just using one drug for cancer treatment, it is usual to combine several drugs in multiple-arm studies, often standard chemotherapy with targeted or/and immunotherapeutic agents. Randomization between the arms in such studies is protection against bias because randomization ensures that characteristics of the study populations between the arms are more likely to be equal than without randomization. (Rubinstein et al. 2005) In a classical randomized setting, as conducted in a phase III study, there is an experimental arm containing the drug candidate, in which patients receive the drug candidate, and a prospective control arm, in which patients receive the standard therapy. These two arms are compared with statistical analysis to evaluate the superiority of the drug candidate compared to the standard therapy. (Dignam et al 2006). Such studies need up to a four times larger sample size compared to a single arm. This large sample size, which goes along with a high financial burden, can be a hurdle when planning and conducting randomized phase II study. Besides the financial aspects, it may be difficult to enroll such many patients, especially when testing a drug of a rare tumor disease. To address these difficulties, phase II randomized studies are usually not designed as phase III randomized studies. Instead, the design and statistical analysis are adapted in a way, which allows a moderate sample size while maintaining an appropriate power. The preferred endpoints of a randomized phase II studies are PFS and OS. PFS is seen as superior in most cases due to certain properties and advantages of PFS, see **section 2.1**. Randomized phase II designs can be grouped into four sections: randomized design with the reference control arm, selection designs, screening designs and discontinuation designs. (Rubenstein 2014)

Randomization design with reference control arm

Herson and Carter proposed a randomized design including a control arm with standard therapy and several experimental arms, which reduce the required sample size. Because of the small sample size, there is no direct statistical comparison between the control arm with the standard therapy and one or several experimental arms. Instead, the efficacy of the experimental arms is compared to a historical control. The control arm evaluates whether the patient population of the treatment arm is comparable to the patient population of the historical control by comparing the outcome of the control arm with the outcome of the historical control. (Herson and Carter 1986) If the outcome of the control arm is significantly better than the outcome of the historical control, the historical control may not be suitable for a comparison with the outcome of the experimental arm(s). A disadvantage is, that the control arm is actually too small to gain sufficient reliable data, whether the historical control is compatible with the control arm and therefore with the treatment arm. (Rubinstein 2014)

Randomized selection design

Randomized selection designs are also described as “pick the winner” designs. In this setting patients are randomized to two or more arms, which contain several drug candidates or different doses and schedules of a drug candidate or different combinations of several drug candidates. Among these different drug candidates, the superior drug candidate will be chosen and tested in further studies. This randomized selection design is a selection tool for prioritizing between different drug candidates, but it is not suitable for a setting in which one or more experimental arm(s) are compared to a standard-treatment control arm. (Rubinstein et al. 2005)

Such a randomized selection design is proposed by Simon et al. Here the most efficacious drug candidate is chosen by the superior response rate with a power of 90% if the true difference between the tested drug candidates is at least 15%. (Simon et al. 1985) A disadvantage of this design is, that there is no comparison with standard therapy, meaning that the superior drug candidate of the “pick the winner” design may be inferior compared to the standard therapy. (Rubinstein et al. 2014)

The statistical design proposed by Liu et al. solved this issue by treating each randomized experimental arm as a single-arm, two-stage design within the same time frame and inclusion criteria, in which the efficacy of every drug candidate is compared to a historical control separately. If a treatment arm shows no sufficient efficacy compared with the historical control, this arm will be dropped. (Liu et al. 2012) This design is often used in practice instead of the design of Simon et al. Of course, this design has some

limitations concerning the use of a historical control, as discussed in the previous **section 2.4.** (Rubenstein 2014)

The strength of the selection design is that only a relatively small sample size is required. As an example, for detecting a 15% difference in the efficacy between two drug candidates with a power of 90%, a sample size of only 29 – 30 is required to meet these conditions according to Simon et al. (Simon et al. 1985)

The selection design is appropriate for selecting among different doses or schedules of the same drug candidate and is therefore suitable to find the appropriate dose with the highest efficacy and lowest toxicity rate. Suitable historical controls should be available to evaluate, whether the drug candidate shows inferior efficacy compared to the standard treatment. (Rubinstein et al. 2005)

Screening design / Comparison design

Screening designs test if the drug candidate is more promising than the standard treatment. For this reason, screening designs are used, if there is no reliable historic control available for comparison. In contrast to the selection design and discontinuation design, a formal statistical analysis between the treatment arms and the control arm is conducted. (Rubinstein et al. 2005) Rubinstein et al. proposed a randomized screening design, in which patients are randomized either to an experimental arm containing the drug candidate + the standard therapy or to a control arm containing the standard therapy alone. This screening design is designed in a way, that statistical properties and sample size meet the criteria of a phase II study. (Rubinstein 2014)

Randomized screening designs are appropriate for targeted and immunotherapeutic drugs. Because these drugs do not aim at tumor size reduction and therefore do not use ORR as primary endpoint, PFS or OS are used for which historical control are not available. (Rubinstein et al. 2014)

Discontinuation design

The purpose of randomized discontinuation designs is to detect if there is a homogenous subgroup, which benefits from the treatment. Only the patients, which are part of the homogenous subgroup, are randomized to the treatment arms or control arm with PFS as primary endpoint. (Ang et al. 2010)

Rosner et al. proposed a discontinuation design, which intends to examine the efficacy of drug candidates in a homogenous group as visualized in ***Figure2***. For a given period, all enrolled patients receive the drug candidate. After an interim evaluation patient with

at least a stable disease are randomized to the group, that continues with receiving the drug candidate or to the group, that receive standard therapy. (Rosner et al. 2002) In a homogenous subgroup, it is more likely to see a larger effect compared to a heterogenous subgroup. With this design, a subgroup, that benefits from this drug candidate can be detected, but on the other hand, it may be difficult to define the characteristics of this subgroup in an appropriate way for further studies. An ethical disadvantage is the large sample size needed to conduct this design. Because only the initial drug candidate responders will be randomized resulting in a large initial sample size. (Rubinstein 2014)

In practice, this design is appropriate, when testing a targeted agent with cytostatic activity with low ORR. A homogenous subgroup can help to differ between slow tumor growth due to the efficacy of the drug or general slow tumor growth. Furthermore, carry-over effects may be a concern in patients, who are initially treated with the drugs candidate and are then randomized to the control arm receiving the standard therapy. (Ang et al. 2010)

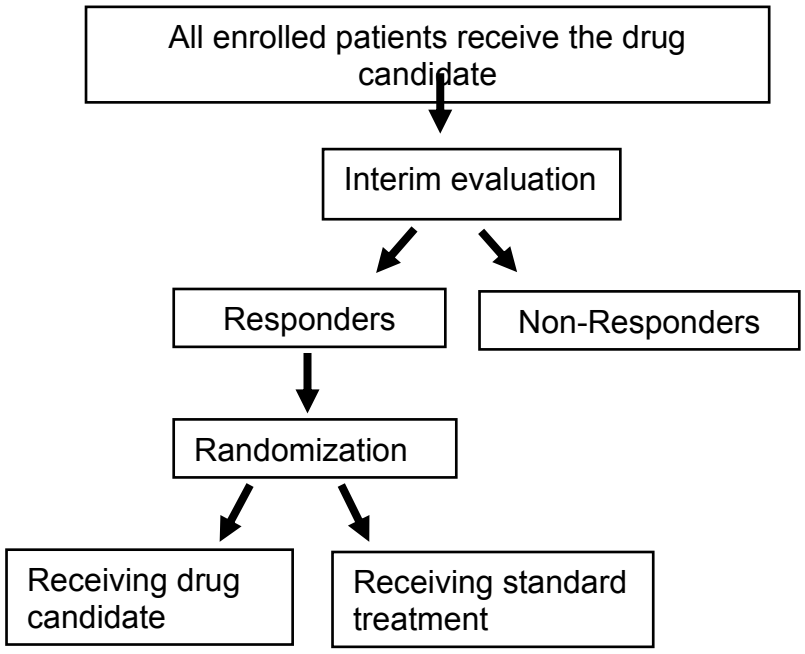

Figure2:Randomized discontinuation design: The aim of receiving a homogenous subgroup is done by randomizing only responders of the drug candidate to either the treatment arm or the control arm with standard treatment. With this design, a certain subgroup responding to the drug candidate can be detected. (Rubinstein 2014)

Advantages and disadvantages of randomized phase II studies

Randomized phase II studies are often criticized for being underpowered. This leads to a high rate of false-negative results. A solution is to increase the type I error level leading to an increased false-positive rate. But there is little concern paid to this high false positive rate because of the following reasons: Randomized studies, especially selection designs, are seen as selection and prioritization tools rather than as “proof of concept” studies. Furthermore, the decision, whether a drug candidate will be tested in a phase III study, depends not only on the result of one single phase II study but on several phase II studies, so a randomized phase II study does not need to fulfill this “proof of concept” paradigm. (Farley and Rose 2010; Ratain and Sargent 2009, Sargent and Taylor 2009)
A single-arm study is appropriate for testing only a few agents for proof of concept and the biological efficacy or for evaluating, whether a drug candidate is superior to the standard therapy for phase III testing. High reliable historical controls must be available to guarantee a reliable comparison to the standard therapy. Single-arm studies are appropriate for tumor diseases and drug candidates, in which the desired outcome will not occur in the absence of the drug candidate. (Rubinstein 2014; Ratain and Sargent 2010)
Randomized phase II designs are the appropriate choice if either many drug candidates are available for testing or for finding the optimal dose of a single drug candidate in one study. Here, randomized studies can be seen as a selection and prioritization tool. They are also appropriate for targeted and immunotherapeutic drugs or for a combination of chemotherapeutic and targeted/immunotherapeutic drugs because there is no reliable historical control for such drug candidates. The same applies to novel endpoints, e.g., time-related endpoints like PFS or biomarkers, see ***Table 3***

2.7 Adaptive Design

A study is conducted as follows: first, the study is planned with an appropriate design, sample size calculation, form of statistical analysis, and so on. After that, the study is conducted with these previous set properties. In the last step, the obtained data are analyzed in the way it was decided in step 1. Generally, all steps and decisions of a study are set in the initial planning phase, after that, changes in design and analysis are not possible to prevent data-driven bias and a decrease in the power. On the other hand, this inflexible setting permits no options for necessary changes during the study conduction. (Pallmann et al. 2018)

Table 3: Recommendation for the appropriate use of single-arm and randomization studies. (Rubinstein 2014, Rubinstein et al. 2005)

Design	Appropriate for
Single-arm design	Testing few drug candidates for efficacy (“proof of concept”) Testing for inferiority if reliable historical control is available Drug candidates, in which the desired outcome will not occur in the absence of the drug
Randomization design with reference control arm	Evaluation, if historical control is suitable for a comparison with the tested drug candidate
Randomized selection design	“pick the winner” design to select or prioritize one out of several drug candidates determine the optimal dose of a certain drug candidate high number of testing drugs available demand of an reliable historical control
Randomized screening design	For testing targeted/immunotherapeutic drugs or combination of chemotherapeutic and targeted/immunotherapeutic drugs No need of a historical control Comparison with the standard-therapy control arm
Randomized discontinuation design	For detecting an unknown subgroup benefiting from the drug candidate

Adaptive designs can overcome these limitations by allowing adaptations to the design and statistical analysis after the beginning of the study. Based on the obtained data of the interim evaluation, decisions about changes in the design or statistical analysis can be done, if necessary. Adaptive designs enable these modifications by keeping the validity and integrity of the study. (Chow and Chang 2008)

Such adaptive designs can include the following changes: a redefinition of the sample size, dropping treatment arms in a multi-arm setting due to inferiority of the drug, changing patients´ assignment to another experimental or the control arm, identifying a sub-population within the sample size, that may benefit the most and drop the other patients or stopping the study early due to inferiority or superiority of the drug candidate. (Pallmann et al. 2018)

Common examples of adaptive designs which include these changes are listed in **Table 4**. Population enrichment designs include the randomized discontinuation designs, which were described in the previous section 2.6. Seamless phase I/II and phase II/II designs

break up the rigor sequence of phase I, phase II, phase II by uniting phases into a single study. Se**ction 2.8.** provides more information about this study design.
In clinical practice, common adaption designs include early stopping rules due to inferiority or superiority of the drug candidate, “pick the winner” or “drop the looser” designs, and adaptive seamless designs. (Chow and Chang 2008)

Table 4:An overview of common adaptive designs and its use (Adapted by Pallmann et al.)

Design	Adaption
Group sequential	Early stopping for safety, futility or efficacy
Sample size re-estimation	Adapting sample size to achieve the desired power
Multi-arm-multi-stage	Options to drop inferior treatment arms or select the winner in interim analysis
Population enrichment	Identifying subpopulation, which benefit the most and dropping all other enrolled patients
Biomarker-adaptive	Using information from biomarkers or adapt on biomarkers
Seamless phase I/II	Combination of safety and efficacy into one study
Seamless phase II/III	Combination of selection and confirmatory stages into one study

Advantages and disadvantages of adaptive designs

Concerning the issue of financial and patients´ resources, an adaptive design can shorten the period, in which the study is conducted as well as reduce the required sample size with ensuring a high chance of reliable results. With this goes along, that fewer patients may be randomized to a treatment arm, that shows no or less efficacy. Adaptive designs can prevent underpowered studies. Bias because of a heterogenous study population can be prevented by population enriched designs. Because of the fact, that some adaptive designs aim at stopping early after an interim evaluation and additionally can adapt the number of stages when the drug shows inferiority or superiority, drug approval is accelerated. To sum up, adaptive designs give the possibility to react flexibly to newly available information and demands occurring in the ongoing study. If there are uncertainties in dose, effect sizes, its variability, and clinical endpoints, adaptive designs are superior. This results in an increase in efficiency. (Pallmann et al. 2018; Lopéz et al. 2012)
A concern of adaptive designs is that the adaptions made in ongoing studies may introduce bias and lead to a high false-positive error rate. Furthermore, the outcome may

be difficult to interpret. Some adaptive designs contain a sample size re-estimation after an outcome is observed, which endangers the results being noisy and prone to bias. But adaptive designs are conducted in such a way, to prevent these concerns. (Lopéz et al. 2012)

Disadvantages of adaptive designs are, that the planning needs much effort and can only be calculated with complex and computationally intensive methods. Because of this, mostly specialized software is needed. It is much more laborious to understand its statistical analysis than the analysis of Gehan´s design (Gehan 1961) or Simon´s Minmax design (Simon 1989). Because of the rare use of adaptive designs, there is no regulatory guidance available. Because of the above-mentioned concerns regarding adaptive designs, it is important to require high transparency, especially for decision procedures. (Lopéz et al. 2012)

2.8 Phase I / II and Phase II / III

Combined studies are studies which unites a phase I and a phase II or a phase II and a phase III study in a single study. These designs are gaining more popularity in the recent years, especially if the activity of a drug candidate is evaluated in combination with known active agents or in a combination with other drug candidates (Wang et al. 2012) This section describes the main idea behind these combined studies inclusive a few popular design proposals and its advantages and disadvantages.

Combined phase I/II

A phase I study is conducted to detect the maximal tolerated dose (MTD) of a drug candidate by detecting the dose-limiting toxicity (DLT). A combined phase I/II study includes therefore the evaluation of the DLT as well as the efficacy of the drug candidate. The general assumption, that the probability of toxicity and the probability of efficacy increases monotonically with increasing dose is not always met in oncological drug candidates. Therefore, a combined, seamless phase I/II study might be superior compared with the sequential phase I and phase II design for defining the optimal efficacy and toxicity balance. Till now, such combined phase I/II studies are not common in practice and only a few designs have been proposed yet. (Wages et al. 2014)

Huang et al. proposed a combined phase I/II design, that identify a set of combination with acceptable toxicity in a “3 + 3” dose-finding design. After that, a phase II is conducted, in which patients are randomized to different experimental arms, that contain the dose levels of the set of combinations of the phase I study. (Huang et al. 2007)

Yin and Yuan proposed a phase I/II design in which doses of acceptable toxicity are established with copula-type regression. For comparing the efficacy of these toxicity-acceptable doses in a phase II study, randomization between these doses-arms is done using a novel procedure, called “moving-reference adaptive randomization”. (Yuan and Yin 2011

Combined phase II/III

Combined phase II/III study designs include a randomized phase II study with a control and one or several experimental arm(s) and a phase III study for confirmatory comparison if the drug candidate shows superiority against the standard treatment. Patients, which are enrolled in the phase II study, will be continued after successful passing to the phase III study. Furthermore, additional patients are enrolled in the phase III study. The phase II part can incorporate early stopping rules. This is appropriate for a drug candidate, for which the superiority is high probably and the phase III part acts only as a check. (Ang et al. 2010)

There is wide variability in the design of combined phase II/III studies including same or different endpoints for phase II and III, variations in the number of experimental arms, and number of interim evaluations. (Wang et al.2012)

Inoue et al. (2002) proposed a seamless phase II/III phase with the two endpoints survival rates and response rates, and early stopping rules in which patients are randomized either to the treatment arm or to the standard arm. Statistical analysis is done with Bayesian methods. (Inoue et al. 2002)

Storer (1990) proposed a design, in which a randomized phase II study with an experimental and standard-control arm is conducted. At the end of phase II, the outcome of the experimental arm is compared with a historical control. Depending on this result, this phase II study is transferred to a phase III study. The control arm is not involved in statistical analysis of the phase II part, but only in the phase III part. (Storer 1990)

Ellenberg and Eisenberger proposed a randomized phase II/III design, which includes a direct comparison between the experimental arm and control arm. As long as the response rate of the treatment arm exceeds the response rate of the control arm, the phase III study continues. The false-negative error rate of the phase II studies is 0.05. The sample size of the phase II part is approximately the double size of a single arm-phase II study with the same targeted difference. (Ellenberg and Eisenberger 1985)

Advantages and disadvantages

Combined phase II/III studies save time, financial- and patients resources because patients enrolled in a phase II study can be enrolled seamlessly to the phase III study.
A concern is the high false-positive error rate in the phase II part, which is necessary to maintain an adequate power in the phase III study. If the outcome of the phase II part is not available after a short period, the study may need to stop temporarily till the desired outcome is available. The sample size required for the phase II part in a combined study is usually larger compared with a phase II study. But because of the enrollment of patients, which are in phase II, to phase III, the sum of the sample size of phase II and phase III studies is smaller than the sum of sample sizes of single-phase II and phase III studies. The infrastructure and conditions for a phase III phase must be developed, even if the phase III study is not conducted due to the inferiority of the drug in the phase II part. (Ang et al. 2010)

3. Description of commonly used Phase II Designs

This section describes the commonly used designs in detail with a focus on their statistical analysis, sample size estimation, and decision-making. The following designs using the frequentist inference, **see 2.2.** The precondition for the use of the frequentist design is a binary endpoint. If time-related endpoints are used, it is common to dichotomize these endpoints by analyzing the survival probability after a certain period. Other commonly used possibilities for dichotomizing are to analyze the median survival probability or survival hazard ratios.

3.1 Fleming´s single-stage design

In this single-stage design, a predefined number of patients is enrolled in the study. After a set period, statistical analysis is conducted to evaluate the efficacy of the drug candidate. The study can be stopped due to efficacy or inefficacy of the drug candidate. For calculation of the required sample size N, that ensures an adequate power of the study, the values of π_0, π_1, the false-positive error rate α, and the false-negative error rate β must be defined.

Let S be the number of patients, who respond to the drug candidate and $\mathrm{r} = \mathrm{S}/\mathrm{N}$ be the response rate following a binomial distribution $\mathrm{B}\ (\mathrm{N}, \pi)$. Let $z_{1-\alpha}$ und $z_{1-\beta}$ be the standard normal deviates of α and β. The Null-Hypotheses is rejected, if the following condition is met:

$$S \geq 1 + Np_0 + z_{1-\alpha}\sqrt{Np_0(1 - p_0)} + 1$$

The additional "+1" preserves the test to be too anti-conservative.

Resulting from this formula, sample size calculation is done as follows:

$$N = \frac{\left\{z_{1-\alpha}\sqrt{[p_0(1 - p_0)]} + z_{1-\beta}\sqrt{[p_1(1 - p_1)]}\right\}^2}{(p_1 - p_0)^2}$$

Sample size calculation is done by approximating the binomial distribution to a standardized normal distribution. This approximation may be imprecise, especially, if $N\pi < 10$. (Fleming 1982)

A´Hern adapts the sample size calculation of Fleming´s design by using the exact Binominal probabilities:

$$C_{A´Hern} = N_{A´Hern} \times (p_0 + (\frac{z_{1-\alpha}}{z_{1-\alpha}++z_{1-\beta}}) \times (p_1 - p_0)$$

3.2 Fleming´s two- or multistage design

Fleming´s two- or multistage design is an extension of his proposed single-stage design. This design allows early termination due to efficacy or inefficacy of the drug candidate; an estimation of the success rate is possible after the first stage. (Kramar et al. 1996)
Let n_i be the sample size of the i^{th} stage with $i = 1, \ldots, g$. Let r_i be the response of the i^{th} stage with $\mathrm{i} = 1, \ldots, \mathrm{g}$.. Let s_i be minimum response rate for the i^{th} stage for which the drug candidate is declared as not sufficient efficacious and a_i be minimum response rate for the i^{th} stage for which the drug candidate is declared as sufficient efficacious. The decision for moving to the next stage or stopping the study is done as follows:

- $\sum_{i=1}^{g} r_i \leq s_i$ stop study due to inefficacy of the drug candidate
- $\sum_{i=1}^{g} r_i \geq a_i$ stop study due to efficacy of the drug candidate
- $s_i < \sum_{i=1}^{g} r_i < a_i$ continue to the next stage

The sample size for the first step is calculated equal to sample size calculation of a single-stage study with controlling the overall false-negative and false-positive error rate for a predefined minimum response rate a_1, for with the drug candidate is declared as sufficiently efficacious and the maximum response rate for which the drug candidate is declared as not sufficient efficacious s_1. The calculated sample size is divided arbitrarily and equally to the stages. For every stage, cut-off points for every stage are calculated as follows:

$$a_i = \left(\sum_{i=1}^{g} n_i \pi_0 + Z_{1-\alpha} \sqrt{N \pi_0 (1 - \pi_0)} \right) + 1$$

$$s_i = \left(\sum_{i=1}^{g} n_i \pi_1 + Z_{1-\alpha} \sqrt{N \pi_0 (1 - \pi_1)} \right)$$

(Fleming et al. 1982)

3.3 Gehan´s two-stage design

Gehan first proposed 1961 a two-stage design, which determines the required sample size for the first stage for a set false-negative error rate. This design allows early stopping after the first stage, if no success in any patients can be observed. If at least one treatment success is observed, this study moves on to the second stage. After this stage, the success rate can be estimated with a predefined standard error. (Gehan 1961)

Let n_1 be the sample size of the first stage. For the first stage the minimum response rate, for which the drug candidate is declared as sufficient efficacious π_1 is set to a specific value. With this value, the chance of observing zero successes among n_1 treated patients given the true response rate of the drug π_1, which is the false-negative error rate β, is calculated as follows:

$$\beta = P(0\ successes\ among\ n_1\ patients\ given\ \pi_1)$$
$$= (1 - \pi_1)^{n_1}$$

By setting β to a specific value, the above-mentioned formula can be resolved for sample size calculation:

$$n_1 = \frac{log(\beta)}{log(1 - \pi_1)}$$

Let r_1 the number of responders observed in the first stage and r_2 the number of responders in the second stage and $R = r_1 + r_2$ the cumulative number of responders of the first and second stage If $r_1 > 0$, the study moves to the second stage with the enrollment of additional n_2 patients, resulting in a total sample size of $N = n_1 + n_2$. The value of n_2 depends on the predefined standard error $SE(R)$ and is calculated as follows:

$$SE(R) = \sqrt{\frac{r_1(1 - r_1)}{n_1 + n_2}}$$

By resolving this equation to N, the formula for the sample size calculation is:

$$n_2 = \frac{r_1(1 - r_1)}{SE(R)} - n_1$$

The underlying assumption of the formula SE (R) is, that the standard error of the response rate of the first stage SE(R) calculated by:

$$SE(R) = \sqrt{\frac{r_1(1 - r_1)}{n_1}}$$

SE(R) is approximately the same as for the response rate of the second stage.

In practice, R cannot be calculated after stage 1, because the values of r_2 and n_2 are unknown. Calculation of n_2 only with the success rate of the first stage r_1 is not reliable and imprecise because of the small value of r_1. For calculating n_2, r_1 is estimated with π_U by using the one-side upper 75% confidence limit of the cumulative Binominal distribution for π:

$$B(r_1; \pi_U, n_{1)} = 0.25$$

Now the formular for n_2 is as follows:

$$n_2 = \frac{\pi_U(1 - \pi_U)}{SE(p)} - n_1$$

(Gehan 1961)

Gehan´s design differs from other frequentist designs by only controlling the false-negative error rate β for rejecting an ineffective treatment for sample size calculation. In this design, p_0, the value of a minimum efficacy is not specified, and therefore, it is not necessary to specify the false-positive error level.

Gehan´s design is easy to understand, implement, and to calculate, however, the required sample size tends to be larger compared to other two-stage designs. (Kramar et al. 1996)

3.4 Simon´s Optimum design and Minimax design

These two-stage designs are an extension of Flemings multistage design by optimizing the required number of patients under the Null-Hypothesis. Early termination after the first stage is only possible due to the inefficacy of the drug candidate. (Simon 1989) Whereas the Optimum design minimizes the average sample size, the Minimax design minimizes the maximum sample size. Early termination due to the efficacy of the drug candidate is not permitted to gain additional and more precise information about the response rate of the drug candidate. (Kramar et al. 1996)

π_0 and π_1 are set to predefined values. Let r_1 be the number of successes in the first stage, r_2 the number of successes in the second stage, n_1 the sample size of the first stage, and n_2 the sample size of the second stage resulting in the total sample size $N = n_1 + n_2$. Let s_1 be the maximum response rate for which the drug candidate is declared as not sufficiently efficacious.
n_1 patents are enrolled in the first stage with r_1 successes observed at the end of the first stage. Let s_1 be the maximum response rate for which the drug candidate is declared as not sufficient efficacious after stage 1 and S the maximum response rate for which the drug candidate is declared as not sufficient efficacious at the end of the study.
Let PET (probability for early termination) be the probability, that the study will be stopped after the first stage if $r_1 \leq s_1$. It is calculated by using the cumulative binomial distribution:

$$PET = B(s_1; \pi, n_1)$$

By setting PET to a predefined value and based on this formula, the expected sample size (EN) required for the whole study with a set power is calculated as follows:

$$EN = n_1 + (1 - PET)n_2$$

After the second stage, the probability of rejecting a drug candidate, if R ≤ S. is:

$$B(s_1; \pi, n_1) + \sum_{x=s_1+1}^{min(n_1,S)} b(x; \pi, n_1)B(S - x; \pi, n_2)$$

Where b denotes the binomial probability mass function. (Simon et al. 1989)

Optimal design

The optimal design minimized the number of patients enrolled in the first phase. Values for π_0, π_1, the false-positive error rate α, and the false negative error rate β are set. Then, values of R_1 and R_2, which met the criteria of π_0, π_1, α and β are calculated computational, for each potential sample size N and n_1. These computationally calculated values for R_1 and R_2, whose corresponding sample size is equal to the calculated expected sample size, are declared as optimal. (Simon et al. 1989)

Minimax Design

The procedure to detect the smallest sample size in the Minimax design is almost equal to the Optimal design expect, that in this setting, values of R_1, R_2 and n_1, and n_2 are calculated for each potential sample size N.
(Simon et al. 1989)

4. Systematic Review: Phase II Designs used in Practice

In the previous chapters, possible group configurations and different methods for statistical analysis have been presented. There are many different possibilities to design and analyze a phase II study. To evaluate, which of these above-mentioned group configurations and statistical designs are currently used in practice, a literature review was conducted. In this chapter, its methods for the selection of suitable studies are described. Furthermore, its results are described including median sample size and median study duration, group configuration, the usage of endpoints, and statistical designs.

4.1 Methods

Paper selection

Only phase I/II, phase II, and phase II/III studies, which test drug candidates and have cancer as subject, were incorporated in this literature review. The search was limited to the following journals: The Lancet (LAN), Lancet Oncology (LO), New England Journal of Medicine (NEJM), and Journal of Clinical Oncology (JCO). Publications, which do not examine drug candidates but the efficacy of surgery interventions, radiation therapy, and tumor treating fields, were excluded.

The search for phase I/II, phase II and phase II/III studies was conducted on the database of the respective journal and on PubMed. The reason for the additional search on Pubmed was done to examine if the search results of these different databases are the same or if there are differences in the search result. Furthermore, this additional search on Pubmed ensures not to oversee any studies. This search on Pubmed was conducted for every journal separately.

Additionally, a search for phase II studies was done on EudraCT (European Union Drug Regulating Authorities Clinical Studies Database), where all authorized studies in the European Union on medical products and drugs are registered. Publications in journals are often affected by publication bias, in other words, studies with promising results are more likely to be published than studies without promising results. In contrast, the EudraCT database is not affected by publication bias, because all European studies must be registered mandatorily. The purpose of including EudraCT in this literature review was to examine, whether there are any differences in group configurations between studies found on the database of journals and registered studies in EudraCT

which may be due to publication bias. Statistical designs could not be compared, because EudraCT provides no information about the statistical design.
The search was done for every journal as follows:

Table 5: Search term and Filters used for study extraction. The rows describe the used filters. The column "Journal" refers to the search on the database of the journal

	Journal´s database	**Pubmed**	**EudraCT**
Search term	cancer OR myeloma OR lymphoma OR glioma OR blastoma OR melanoma OR tumor OR sarcoma OR carcinoma AND phase 2 OR phase II AND [+ journal]	cancer AND phase II OR phase 2 AND [+ journal]	cancer OR myeloma OR lymphoma OR glioma OR blastoma OR melanoma OR tumor OR sarcoma OR carcinoma AND phase 2 OR phase II
Article type	Research article (LAN, LO, NEM,	Clinical Study, Phase II	Study Phase II
Year	2019 (LAN, LO) 2019/2020 (NEM)	2019 (LAN, LO) 2019 / 2020 (NEM) Jan - Jul 2020 (JCO)	2020
Subject	-	Cancer	
Journal	The Lancet (LAN) The Lancet Oncology (LO)	-	
Specialty	Haematology/Oncology (NEM)	-	

For The Lancet, the search was limited to the year 2019, because the most recent studies should be evaluated and the university license for unlimited access is only available up to the year 2019. On the database of the journal, the search was done with the following search term: "cancer OR myeloma OR lymphoma OR glioma OR blastoma OR melanoma OR tumor OR sarcoma OR carcinoma AND phase 2 OR phase II AND Lancet". Following additional filters were set: journal: The Lancet; Publication date: 2019; article type: research article. On Pubmed, the search for this journal was done with the

following search term: “cancer AND phase II OR phase 2 AND Lancet.” and following filters: publication date: 2019; article type: clinical study phase II; subject: cancer.
The search term and filters for The Lancet Oncology were the same as for The Lancet, except the filter journal, which was set to “The Lancet Oncology” instead of “The Lancet”.
The search term on Pubmed was “cancer AND phase II OR phase 2 AND Lancet Oncol.” with the same filters as for The Lancet. With this search term, not only publications of LAN are suggested, but publications of all journals belonging to the Lancet family. Only studies published in LAN are selected from the search result.
The search for New England Journal of Medicine was limited to the years 2019/2020, because university license for this journal was available till 2020, in contrast to The Lancet and The Lancet Oncology. For the search on the database of the journal, the search term was the same as for the other journals. Following filters were set: date: 2019-2020; article category: research article; specialty: Hematology/Oncology. The search term for the search on Pubmed was: “cancer AND phase II OR phase 2 AND N Engl J Med.” With filters publication date: 2019-2020, article type: Clinical Study, Phase II and subject: cancer.
The search for Journal of Clinical Oncology was limited to January-July 2020. This period was chosen because the university license for this journal is only available till June 2020. Because of the large number of published phase II studies in 2019 and limited time for the research, 2019 was excluded. The search for this journal was only done on Pubmed, because on the database of the journal, filtering for a certain time is not possible. “Cancer AND phase 2 OR Phase 2 OR phase II OR Phase II AND J Clin Oncol” was used as the search term. Following filters were set: publication date: from 2020/01/02 to 2020/12/31; article type: Clinical Study, Phase II, subject: cancer.
Phase II studies in EudraCT were extracted with the following search term: “cancer OR myeloma OR lymphoma OR glioma OR blastoma OR melanoma OR tumor OR sarcoma OR carcinoma””. Following additional filters were set: study Phase: Phase II, date Range: 2020/01/01 to 2020/12/31, study status: completed. **Table 5** provides an overview of the different search terms and applied filters.
The search term in EudraCT differs from the search term used on the database of the relevant journal because in Pubmed, there is a filter for selecting cancer as subject. The database of the relevant journal does not provide such a filter. Studies with cancer do often not contain the term “cancer”, but the term for the special tumor type. Because of this, the search term for the search on the database of the journal contains these terms.

Data extraction

Following data were extracted from papers based on the search on the journal´s database and Pubmed: phase (I/II, II or II/III), tumor type, sample size, multicenter, open-label or controlled, usage of placebo, randomization, primary and secondary endpoints, number of arms and stages, statistical design, planned power, one-sided or two-sided type I error-level, duration of enrollment, the aim of the study and which reference is used to evaluate efficacy of the drug candidate.
Information about statistical analysis of the registered studies in EudraCT is not given and instead of enrollment duration, information about the expected duration of the study is provided. In contrast to the journal´s publication, EudraCT provides information, whether the study contains parallel arms.

Statistical methods

The aim of the statistical analysis was a description about the practical used group configuration, design, and statistical analysis. Outcomes from publications extracted from the database of the journal/Pubmed and from EudraCT were analyzed and reported separately to detect any differences that may be due to publication bias. No separate analyses between different years or different journals were made because of the small number of publications per journal and year. Nominal variables e.g., primary, and secondary endpoints or statistical designs were reported as a rate in percentage and as a cumulative absolute value. Numeric variables e.g., the sample size of the study or study duration were reported by calculating median, minimum and maximum. Because of the great variance of the sample size median was calculated instead of the arithmetic mean. In EudraCT, there is information given about the estimated study duration in the Member state as well as in all countries in which the study was conducted. Here, information about the estimated study duration in all countries was selected. Analysis was done with R 4.0.5.

4.2 Results

Sample description

The selection process of studies, that met the above-mentioned conditions, is displayed in **Figure 3**. A total number of 126 papers were identified in the database of the journals (LAN: 33, LO: 74, NEJM: 19) But only a few of them met the above-mentioned inclusion criteria. 5 publications listed in the search result of the database of LO were excluded because they are phase I studies. 66 publications (LAN: 32, LO: 24, NEJM: 10) were phase III studies and therefore excluded. 37 publications (LAN:6, LO: 23, NEW: 6) did not evaluate a drug candidate. Instead, they evaluated other interventional therapies against tumors like surgical resection or radiotherapy. Other publications estimated the prevalence of certain tumor types, were meta-analyses or compared certain imaging tools or evaluated the diagnosis of cancer-based on network models/machine learning tools. Drug candidates of 14 publications (LAN: 6, LO: 2, NEJM: 6) did not aim at tumor diseases, but at other diseases like cardiovascular or inflammatory diseases. After this exclusion, a total sample size of 23 publications could be found on the journals´ database. Additional 4 publications in the JCO and 6 publications in NEJM could be found on Pubmed, that were not listed in the search result of the databases of the journals. For JCO, the search was only conducted on Pubmed. Here, 27 studies met the inclusion criteria. Finally, a total number of 60 studies remained, that met the inclusion criteria.

Table 6 provides information about the number of extracted studies from every journal and EudraCT. On EudraCT, a total number of 57 approved phase II studies were found. 14 phase II studies were excluded because their subject was not tumor but other diseases, especially corona virus and inflammatory bowel diseases. After this exclusion, a total number of 43 registered studies remained.

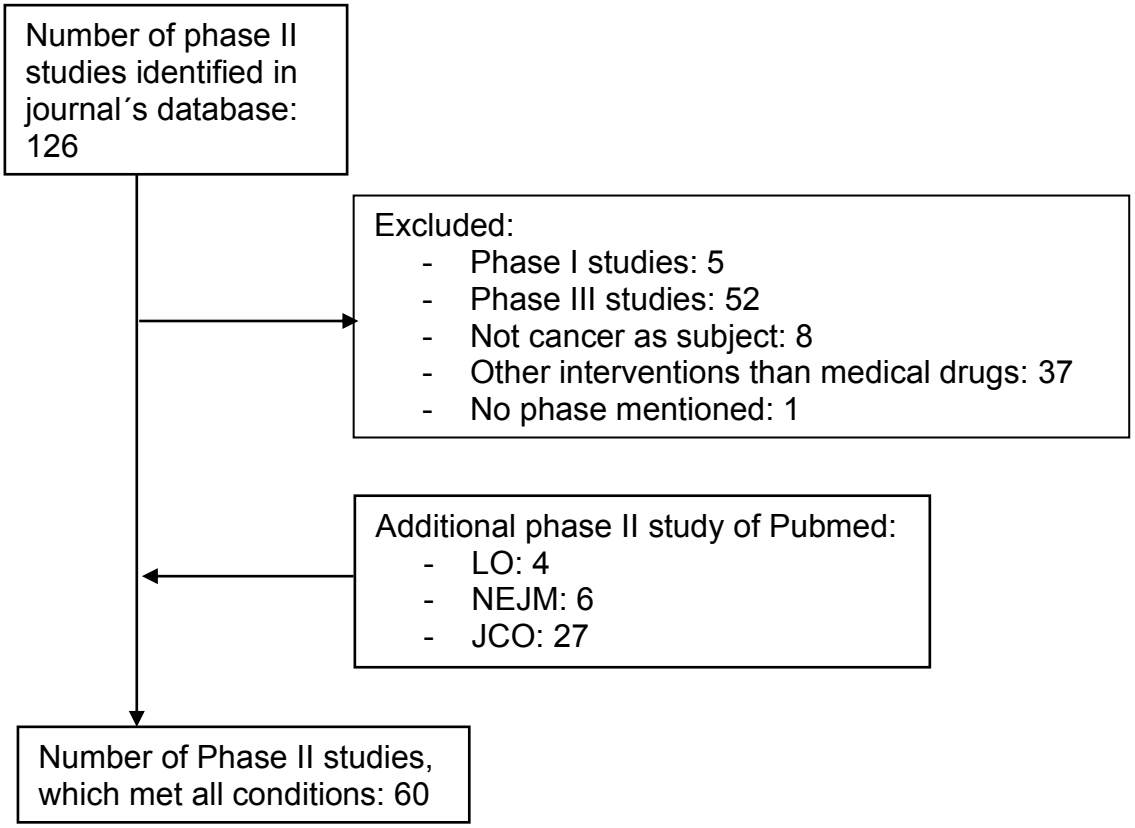

Figure 3: Selection process of publications found on the journal´s database and on Pubmed. Note, that for Journal of clinical oncology, the research was only conducted on PubMed.

Table 6: Number of Publications found in journal´s databases (column name: "Journal") and in EudraCT that met the inclusion criteria

Journal	Journal	Pubmed	Time
Lancet	1	0	2019
Lancet Oncology	20	4	2019
New England Journal of Medicine	2	6	2019/2020
Journal of Clinical Oncology	0	27	01/2020 – 06 2020
EudraCt	43		2020

General Characteristics

In the following sections, studies found on the database of the respective journal/Pubmed are just described as recently published studies, and studies found on EudraCT are described as recently approved.

The selected papers included a wide range of different tumor type, for which drug candidates were tested, see **Appendix Table 11.** The most common tumor types were breast cancer, kidney cancer, stomach/esophagus cancer and immune system-related cancer in recently published studies, and lung cancer, breast cancer, solid tumors cancer, immune system-related cancer and urothelial/bladder cancer in recently approved studies.

In **Table 7**, general characteristics like study duration, the proportion of phase I/II, pure phase II and phase II/III and group configurations are summarized. The most recently published and recently approved studies were pure phase II studies. The portion of recently published phase II studies were with a percentage of 83% higher than the portion of recently approved phase II studies (65%). 15% of all recently published studies and 30% of recently approved studies are phase I/II studies. The smallest part was phase II/III studies. There was only one phase II/III study (2%) among all recently published studies and two (5%) phase II/III studies among all recently approved studies.

The majority of recently published (87%) and all recently approved studies were multicenter, only 13% of recently published studies were single-center studies.

Most studies were designed as single-arm or two-arm studies. Among recently published studies, one half were single-arm and 42% were two-arm studies. Among recently approved studies, the portion of single arm studies was with 58% slightly greater but the portion of two-arm studies was with 31% smaller compared to recently published studies. Studies, which contain more than two arms were very rare: the portion of three-arms studies of recently published studies was 5% and of recently approved studies 7%. 3% of recently published studies and 2% (only one study) of recently approved studies were four-arm studies. One recently approved study was designed as a six-arm study.

All selected recently published studies were either single-stage or two-stage. The majority (73%) were designed as single-stage and 27% were designed as two-stage studies with an interim evaluation. There was no information available about the number of stages of recently approved studies.

Table 7: Design characteristics of recently published and approved studies The absolute number is written in brackets.

	Recently published studies	Recently approved studies
Phases		
Phase I/II	15 % (9)	30% (13)
Phase II	83% (50)	65% (28)
Phase II/III	2% (1)	5% (2)
Multicenter		
Yes	87% (52)	100% (43)
No	13% (8)	0
Number of arms		
1	50% (30)	58% (25)
2	42% (25)	31% (13)
3	5% (3)	7% (3)
4	3 % (2)	2% (1)
6	0	2 % (1)
Number of stages		
1	73% (43)	NA
2	27% (16)	NA
Median study duration (in months)		
Phase I/II	34 (8 -54)	36 (13-86)
Phase II	27 (5 -96)	44(9-72)
Phase II/III	34	54 (48-60)
No information given	18% (11)	2% (1)
Median planned sample size of Median (Min-Max) (		
Phase I/II	75 (30 - 418)	220 (49 - 1500)
Phase II	90 (20 - 532)	120 (42 - 790)
Phase II/III	486 (210 – 761)	620 (600 - 640)
Median planned sample size (Min-Max) per number of arms		
1	46 (20 - 532)	100 (42 - 1500)
2	118 (30 - 761)	166 (56 - 640)
3	102 (89 - 260)	189 (120 - 372)
4	163 (123 - 204)	220
6	0	550
Study success		
Yes	90% (54)	NA
No	10% (6)	NA

For recently published studies, information about the patients´ enrollment duration was provided. In contrast, recently approved provided information about the planned duration of the study. Median duration of enrollment in recently published studies was higher in phase I/II and phase II/III studies compared to phase II studies. The Median duration of recently published phase I/II studies was 34 months, of recently approved phase I/II studies 36 months and of phase II studies 27 months and 44 months respectively. There was only one recently published phase II/III study with an enrollment duration of 34 months. There were only two recently approved phase II/III studies with a median duration of 54 months. 18% (11 studies) of recently published studies and 2% (1 study) of recently approved studies did not provide any information about enrollment duration and expected study duration, respectively. Most of the studies do not report the follow up time.

The median planned sample size of recently published phase I/II study was 75 with a minimum of 30 and a maximum of 418, median planned sample size of recently approved phase I/II studies was higher with a median of 220 (Min: 49, Max: 1500). For recently published phase II studies, the median planned sample size was 90 and for recently approved studies 120. The planned sample size of recently published phase II/III studies was 486 and the median planned sample size of recently approved studies was 620. **Table 7** provides additional information about the median sample size grouped after arms. The more arms were included in one study, the larger the median sample size. The assessable sample size was always equal to or greater than the planned sample size, so the planned power could be held in all reviewed studies.

Only 10% of recently published studies failed, information about success (drug candidate shows sufficient activity and further phase II studies or phase II studies were recommended for this frug candidate) of recently approved studies is not given.

Characteristics of studies including two or more arms

Table 8 considers only studies, which included two or more arms. This table provides information about the portion of randomized, controlled, placebo-used studies. Most of the recently published studies with two or more arms were randomized (97%), only one study (3%) was not randomized. In EudraCT, all studies with two or more arms were randomized.

In 87%, the allocation ratio to the two arms was 1:1 and in 13%, the allocation ratio to the two arms was 2:1. Two studies with an allocation ratio of 2:1 included one arm with the drug candidate and one arm with the standard therapy as control arm. One study included one arm with the drug candidate and one arm with the placebo, because there

was no effective standard therapy for the tumor type, or the tumor type was resistant to chemotherapy. One study also included one arm with the drug candidate and one arm with the placebo. But in contrast to the study mentioned before, patients in the placebo-control arm as well as in the arm with the drug candidate received best supportive care. There was no further specification about this supportive care.

17% of recently published studies and 22% of recently approved studies were double-blinded. All recently published and approves studies, which were double-blind, incorporated a control arm with placebo or standard therapy. There was no blinding between several arms containing different drug candidates. There was no single-blinded study.

Table 8:Characteristics of studies with more than one arm. The absolute number is written in brackets.

Characteristics	Recently published studies	Recently approved studies
Randomized		
Yes	97% (29)	100% (18)
No	3% (1)	0
Randomization Ratio		
1 : 1	87% (26)	NA
2 : 1	13% (4)	NA
Blinded		
No	83% (25)	78% (14)
Single-blind	0	0
Double-blind	17% (5)	22% (4)
Controlled		
Yes	12% (7)	68% (13)
No	88% (53)	26% (5)
No information given	0%	5% (1)
Placebo		
Yes	13% (4)	21 % (4)
No	87% (26)	74% (14)
No information given	0	5% (1)
Parallel arms		
Yes	NA	63% (12)
No	NA	32% (6)
No information given	NA	5% (1)

The majority of recently published studies were open label. With recently approved studies, it was the other way round: the majority (68%) are controlled, for one study (5%), no information about a controlled or open-label design was given.
Only a small number of studies used placebos: The portion of recently published studies incorporating a placebo was 13% and the portion of recently approved studies was 21%. For one recently approved study (5%), no information was given about the usage of placebo. Only in recently approved studies, there was information about the use of parallel arms given. Here, 63% of the studies incorporated parallel arms, for one study, no information was provided.

Endpoints

ORR is defined within a certain period. In most studies, this information about the time was not provided in the statistical part. In all studies, ORR was measured after the RECIST criteria. If time-related endpoints were used as primary endpoints, they were dichotomized by using the median value, e.g., median-PFS or hazard ratios of PFS or OS, or by evaluating the survival rate after a predefined time, often after 3 or 6 months.
The most common primary endpoint was ORR (recently published studies: 56%, recently approved studies: 60%). The second most common primary endpoint was PFS. 16% of recently published studies and 21% of recently approved studies used this time-related endpoint. Only 3 % of recently published studies and 5% of recently approved studies used OS. AE was more commonly used as a secondary endpoint than as a primary endpoint (recently published studies: 3%, recently approved studies: 2%). pCr, minimal residual disease (MRD), disease progression (DP), disease-free survival (DFS), disease control (DC), and complete response (CR) were rare with only a portion equal to or less than 2%, see **Figure 4.**
In recently published studies, all studies used only one primary endpoints except two: one used both pCr and ORR, and another study used three primary endpoints: Adverse effects (AE), ORR, and pCr. One study on EudraCT used OS as well as PFS as primary endpoints. 6 recently published studies used special primary endpoints, labeled as “other” in **Figure 4**. One used the proportion of each group who completed 2 cycles of treatment and initiated the third cycle. Another one defined the rate of changes in the sum of the largest diameter of targeted marker lesions as a primary endpoint. Other special primary endpoints were change in 6 months MRI tumor completion of treatment for feasibility, CR-rate of pola BR vs. BR, time to second PSA progression, and duration of severe neutropenia. Overall, 5% of recently published studies and 11% of recently approved studies used such special primary endpoints.

There were more different secondary endpoints used than primary endpoints per study. In contrast to primary endpoints, where usually only one single endpoint is used, it was common, to use on average four secondary endpoints per study, see **Figure 5**. OS was the most common secondary endpoint (recently published studies: 17%, recently approved studies: 18%).

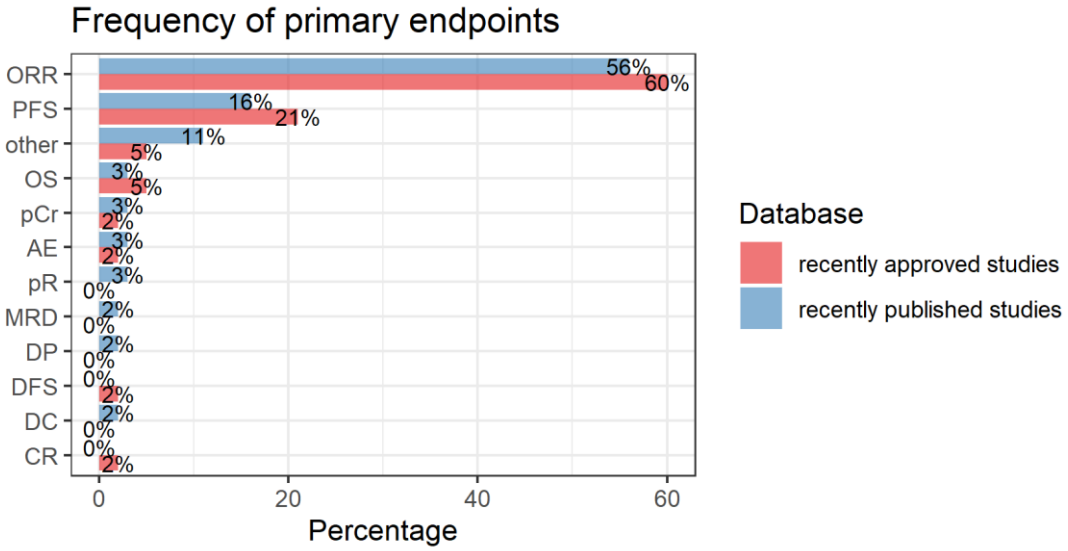

Figure 4: Usages of different primary endpoints in recently published and recently approved studies.

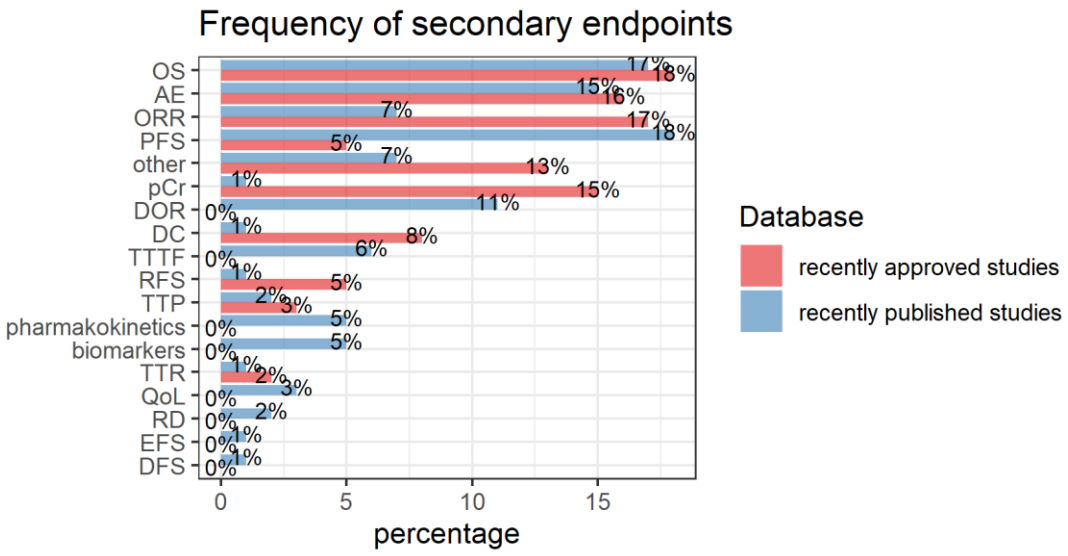

Figure 5: Usage of different secondary endpoints in recently published and recently approved studies

Another commonly used endpoint was the occurrence of AEs. (recently published studies: 15%, recently approved studies: 16%). In general, recently published studies used more different secondary endpoints than recently approved studies. Time to

treatment failure (TTF), biomarkers, pharmacokinetics, QoL, DFS, and event-free survival (EFS) were not used in recently approved studies, but in recently published studies with a portion of 6%, 5%, 5%, 3%, 1%,1%, respectively. ORR as secondary endpoint occurred in recently approved studies with a portion of 17% more frequent than in recently published studies with a portion of only 7%. Only 2% of recently approved studies incorporated time to response (TTR) as secondary endpoint, in recently published studies this portion was 11%. Relapse-free survival (RFS) and TTP were rarely used (recently published studies: 1% and 6%, recently approved studies 5% and 3%). Biomarkers were always used as exploratory secondary endpoints. Some studies used special secondary endpoints. Some examples, which are used in recently published studies were margin-free resection rate, PSA response, time to PSA progression, time to first tumor response, tumor PD-L1 expression. Special endpoints in studies in EudraCT were time to next treatment, minimal residual disease, fatigue, metastatic free survival, and clinical benefit rate. Overall, the portion of special endpoints in recently published studies was 7%, and in recently approved studies 13%

Statistical designs mentioned

67% of all recently published studies did not mention a specific design. One study (2%) used a Fleming single-stage design. 5 studies (8%) used Simon´s Optimal design and 4 studies (7%) used Simon´s Minimax design. 4 studies (7%) mentioned using a Simon´s Two Stage Design without further specification, whether the Optimal or Minimax design is meant. The statistical design of O´Brien Fleming was used in 4 studies (7%). One study reported using the 2 * 2 factorial design for statistical analysis and one study used Randomization with reference arm, see **Table 9.** Half of the studies used a power of 80% and 23% used a power of 90%. The proportion for other power levels was quite small: 70% power was used in 1 study (2%), a power of 85% was used in 2 studies (3%). 2 studies (3%) used 89%, one study used 94% power and two studies (3%) used 95% power. 8 studies (13%) did not provide any information about the power. These studies did not have a formal sample size calculation. Furthermore, two of them did not conduct any hypothesis testing. Instead, they use confidence interval for evaluating the efficacy of the drug candidate. In this case, there was no comparison of the efficacy of the drug candidate with the efficacy of an historical control or standard treatment.
Most of the studies used a type I error level of 0.05 (38% of all studies) or a type-I error of 0.1 (30% of all studies). The portion of studies using a type I error level of 0.025 was 8% and the portion of studies using a 0.2 type error level was 5 %. A type-I error level of

0.15, 0.16, and 0.25was used in one study each. In 13% of all studies, no information about the type I error level was provided.

Table 9: Statistical Designs, power level and type- I error level used in recently published studies.

	Recently published studies
Statistical design	
Single stage	
Fleming single stage	2% (1)
Two stages	
Simon´s Optimal Design	8% (5)
Simon´s Minimax Design	7% (4)
O´Brien-Fleming	7% (4)
Simon´s Two Stage Design	7% (4)
2 *2 factorial design	2% (1)
Randomization with reference arm	2% (1)
No specific design mentioned	67% (40)
Power	
70%	2% (1)
80%	50% (30)
85%	2% (1)
89%	3% (2)
90%	23% (14)
93%	2% (1)
94%	2% (1)
95%	3% (2)
No information given	13% (8)
Type I error level	
0.025	8% (5)
0.05	38% (23)
0.1	30% (18)
0.15	2% (1)
0.16	2% (1)
0.2	5% (3)
0.25	2% (1)
No information given	13% (8)
Type I error-level one or two sided	
One-sided	35 % (21)
Two-sided	15% (9)
No information given	50% (30)

35% used a one-sided and 15% a two-sided type I error level, half of the studies did not provide any information. For studies, which mentioned a statistical design, there was no information provided, if a one-sided or two-sided type I error level was used, except for two studies using a Simon´s Optimal two-stage design with a one-sided type I error level.

There is no relationship between the used group configuration and the use of a one- or two-sided type I error level.

Information about the Null Hypotheses and the expected clinical improvement were provided in the statistical part in all recently published studies. But only in 5 (8%) recently published studies, there was an explanation, why a certain value as expected clinical improvement was chosen. In only 4 (13%) of all recently publishes studies, which used a historic control, information about the origin of the historical control was given. Two third of all studies did not provide information about the test, for which sample size calculation was done. In studies, which reported the test, mostly, a one- or two-sided binomial test was used for binary endpoints and log-rank tests for time-related endpoints. For binomial endpoints, binomial tests, or student´s t tests for the comparison with a historical control and Chi-squared test, Fisher´s exact test or student test for comparing two or more arms and for comparing different subgroups and associations between response and various covariates. For time-related endpoints and for comparing different arms of time-related endpoints lifetime analysis with Kaplan-Meier curves and Hazard Ratios, Cox regression and log rank test for time-related endpoints.

Proof of superiority

Half of the recently published studies and more than a half of recently approved studies used historical controls for testing the drug candidate for superiority (**Table 10**). 70% of recently published studies, which contain more than one arm and 72% of recently approved studies, which contain more than one arm, use a control arm (standard therapy or placebo). A portion of 23 % of recently published studies compared the drug candidate together with the standard therapy vs. the standard therapy alone. Compared with the portion of recently published studies, this portion was with 28% of recently approved studies slightly higher. The portion of studies using a comparative design, which compares more than one drug candidate on superiority, was smaller for both recently published than for recently approved studies (recently published studies: 7%; recently approved studies: 12%). Other ways to compare the efficacy of the drug was only rarely used. The concept of evaluating the drug candidate vs. placebo was used in 5% and drug candidate vs. standard therapy was used in 7% of recently published studies. Two studies compared the drug candidate + standard therapy vs. placebo + standard therapy. The concept of drug candidate + placebo was used in only one single recently approved study. A crossover design and the concept of timing, that means, that the drug candidate was given at different points in time during treatment, was used once each in recently published and recently approved studies. Only one recently published study tested

several drug candidates (in every arm in combination with the standard therapy) for superiority and one study tested for superiority of a drug candidate, which was given on different time schedules. The remaining recently published studies, which contained three or more arms, tested different dose schedules of the same drug for superiority Three recently approved studies tested several drug candidates for superiority, and 2 recently approved studies tested different dose schedules of the same drug candidate for superiority. All recently published and approved studies, which contained two arms, did test the efficacy of just one drug candidate.

To sum up, recently published studies used more different strategies than recently approved

Table 10: Overview in which way the efficacy of the drug candidate is verified for superiority

Proof of superiority/comparative	Recently published studies	Recently approved studies
Historical control	50% (30)	58% (25)
Drug candidate + standard therapy vs. standard therapy	23% (14)	28% (12)
Comparative/superiority	7% (4)	12% (5)
Drug candidate vs. placebo	5% (3)	2% (1)
Drug candidate vs. standard therapy	7% (4)	0%
Drug candidate + standard therapy vs. placebo + standard therapy	3% (3)	0%
Timing	2% (1)	0%
Crossover	2% (1)	0%

5. Discussion

A literature review of phase I/II, phase II, and phase II/III cancer studies in selected peer-reviewed journals as well as in EudraCT was done. The search was limited to the years 2019/2020 in these selected peer-reviewed journals, and to the year 2020 in EudraCT. 60 recently published studies in journals and 43 recently approved studies in EudraCT were reviewed with extracting information about group configuration and statistical designs. The aim of this review was to evaluate differences in theoretically proposed and practically used statistical designs and group configurations.

Although numerous statistical designs have been developed in the last two decades (Hess 2007), the most used design in practice is still an open-label single or two-arm study with a historical control according to this literature review. There is a multitude of proposed theoretical statistical phase II studies designs, that aim on different group configurations, stages, and endpoints, e.g., one-stage or two-stage designs, designs for binary or multinominal outcomes, designs for time-to-event endpoints, decision-theoretic designs, Bayesian inference design, designs for including randomization for naming a few. (Brown et al. 2014). In contrast, in this literature review, only in less than a half studies, a statistical design was even mentioned. In the remaining studies which mentioned a statistical design, a Fleming single-stage design or Simon´s Optimal/Minimax Design and Two-Stage O´Brien-Fleming Design was used. These designs are probably used frequently because they are simple and therefore easy to understand and easy to implement. In contrast, many other theoretically proposed statistical designs for phase II studies are complex and therefore complicated to understand and implement. This may be a reason, why these designs are rarely or not used in practice. A portion of 67% did not mention a specific statistical design. For single-arm studies using a historical control as comparison, a Fleming single-stage design can be assumed based on the information about the hypotheses and sample size calculation. A reason for the poor reporting of this design might be, that the way of sample size calculation in the Fleming´s design is a common way of doing sample size calculation and researcher may not be aware that this is a Fleming design. All two-arm studies with an experimental arm containing the drug candidate and a control arm containing the standard arm did not mention a statistical design. But based on this group configuration a screening design may be used. However, because the description of the statistical procedure did not differ from single-arm studies using a historical as comparison, it is not clear, if sample size calculation and statistical analyses is conducted according to the

screening design or if just the group configuration of a screening design is used but not the statistical analysis of the screening design.

The power of phase II studies is usually set to 80% or 90% (Winter and Pugh 2019, Rubinstein et al. 2009). This is in agreement with the literature review: approximately two-third used a power level of 80% or 90%. But there are also other power levels used within a range of 70% to 95%. The reason for choosing such uncommon power levels is not mentioned in the statistical part of the publications. An explanation for such power levels may be the use of predefined sample sizes due to a restricted study budget. Usually, sample size calculation is done on basis of a predefined power level. But if the sample size is predefined in advance, the power level is a result of this predefined sample size. In some studies, unusual type I error levels were used. There is no explanation in the statistical part given about the reason for choosing these uncommon type I error levels.

In all reviewed studies, the frequentist inference was used, although there are some theoretical statistical designs proposed using a Bayesian design. In practice, however, Bayesian inference methods do not seem widely spread despite some advantages over the frequentist inference, e.g., Bayesian inference enables a more frequent monitoring and interim decision making, see **section 2.3** for more details. A reason for the rare use of the Bayesian inference might be, that statistical designs using Bayesian inference are more complex and therefore not easy to understand implement as the frequentist inference. Furthermore, it is common to use the frequentist inference/hypotheses testing in the medical area.

In most cases, statistical procedures in the statistical part are not precisely described. Although hypotheses are well described, there was poor reporting about the statistical test used for sample size calculation, on which assumption the value of clinical relevance was chosen, and which historical control was used. The statistical part was largely just an enumeration, which statistical tests were used for comparing the efficacy of the drug candidate with the control. There was also poor reporting of the power in case of a predefined sample size.

Thezenas et al and Ivanova et al. did a literature review to evaluate the used group configuration and statistical design of phase II cancer studies published in 2000 and 2014 . In contrast to this literature review, they did not include phase I/II and phase II/III studies in their review. Some changes can be seen in the design and statistical analysis between phase II studies published in 2000, 2014, and this literature review. Some changes include the increased portion of multicenter studies is recent years. Compared to the literature review of Thezenas et al, which was limited to the year 200, the portion of multicenter phase II studies published in 2014 and 2019/2020 has been increased

(Thezenas et al.:43.8%, Ivanova et al.: 91%, my review: 87%). There was also an increased use of ORR as primary endpoint in this review compared to the review of Ivanova et al. (Ivanova et al.:43%, my literature review: 56%). The reporting of hypothesis, power level, and type I error level has been increased compared to the review of Ivanova et al. Whereas in the literature review of Thezenas et al, 46% of all studies published in 2000 provides information about the hypothesis, power and type I error level, the hypothesis formulated in all studies of my literature review and information about the power and type I error level was provided in 87%. The portion of mentioned statistical designs and the type of design used (mainly Fleming´s single-stage design, Simon´s Optimum/Minimax design, and Gehan) was almost similar in all reviews. In agreement with the results of the literature review of Thezenas et al, there is no difference between the planned sample size and the assessable sample size. (Thezenas et al. 2004, Ivanova et al. 2016)

The majority of recently published studies and all recently approved studies, which contain two or more arms, randomized between these arms. In most cases, a randomized setting is chosen, in which the sample size is equally subdivided between the experimental arm with the drug candidates and the control arm containing the standard therapy as comparison.

Two- and more stages designs are highly recommended in literature because there is the possibility of stopping the study earlier due to sufficient efficacy or inefficacy. Despite these advantages of an interim analysis saving time and financial resources, less than a third of the reviewed phase II studies used an interim analysis. In practice, there may be some difficulties to include an interim analysis in the study because this requires an interruption of patients´ enrollment, which goes along with organizational difficulties

Although publications report a decrease in the usage of ORR as primary endpoint, which has been historically the most popular endpoint, in favor of an increasing use of OS and PFS (Kilickap et al. 2018, Wu et al. 2011, Thezenas et al. 2004, Ivanova et al. 2016) more than a half of recently published and approved studies still used ORR as primary endpoint in this literature review. Depending on the drug candidate and purpose of the study, different primary endpoints are recommended, see **section 2.1**. Almost all reviewed studies evaluated targeted or immunotherapeutic drug candidates. For these drugs, the use of time-related endpoints is recommended instead of ORR. There seems to be still a large discrepancy between theoretical recommended endpoints and practically used endpoints.

In contrast to phase II/III studies, which are rarely used, phase I/II are gaining more popularity. This finding goes together with the findings of Wang et al. (Wang et al. 2012)

For this phase I/II studies, there was no specific design for phase I/II studies used. Instead, for the phase I part, a 3 + 3 factorial design was used and for the phase II part, the same designs as for single phase II studies are used. A reason for this could be the fact, that there aren´t many phase I/II designs proposed yet, although phase I/II studies have some advantages compared to single phase I and phase II studies: Combined phase I/II studies save time and financial resources. The sample size needed for a combined phase I/II study is smaller compared to the sum of sample sizes needed in a phase I and phase II study.

Only three recently published studies and 3 recently approved studies distinguished in their description between phase IIa and phase IIb studies. Based on this result, there seems to be just a theoretical distinction between the subtypes of a phase II study, for practical applications, mostly no distinctions are done. This may be due to the fact, that in practice, both aims, collection information about efficacy and safety and identifying and selecting promising drug candidates for a phase III studies, are tested together in one study. This would save financial and patients' resources.

All journals provided the same information on group configuration and statistical analysis. In NEJM, some information about the countries, in which this study takes place, and the final sample size were not in the publication itself but in supplementary data. Furthermore, in contrast to the other journals, the titles of the studies published in NEJM did not contain information about the phase in the title but in the abstract or main part.

There were some differences in the design of studies of recently publishes studies extracted from the journals and recently approved studies extracted from EudraCT. The portion of phase I/II phases was twice as high on recently approved studies compared to the portion of recently published studies. Whereas the portion of recently published studies being multicentric was 87%, all recently approved studies were multicentric. Recently published studies had a slightly lower portion of one-arm and a higher portion of two-arm studies compared with recently approved studies. The median sample size of recently approved studies was higher than recently published studies. Only 12% of recently published studies, but 68% of recently approved studies were controlled, The usage of placebo in recently published studies was lower compared with recently approved studies. Recently published studies used more different primary and secondary endpoints than recently approved studies. These differences can be explained by publication bias, of which recently published studies are affected. Another explanation for this effect is the small sample size of studies in both sources and the study research in only four journals. Another reason for these observed differences is, that recently published studies were limited to an earlier time than recently approved

studies. In recent years, phase I/II studies have become more popular and there is an increased use of randomization. These statements are in agreement with the results of the literature review: the portion of phase I/II studies and randomization was higher in recently approved studies, which were conducted more recent compared to recently published studies.

This literature review has some limitations. First, the sample size of included studies is small. The search for studies is limited to only 4 journals with a high impact factor. So, this literature review does not provide any information about group configuration and statistical design of studies, which are published in journals with a lower impact factor. Additionally, because of the limited research to only 4 journals, only a small portion of all published phase I/II, phase II, and phase II/III studies in 2019/2020 are included in this literature review. Therefore, these results may not display the general state of the group configuration and statistical design of phase II studies in oncology. As mentioned before, the publication bias could be influencing the results. Even the comparison of its results with the results of studies registered in EudraCT could not erase this possible bias, because the recently published studies were planned earlier than the studies on EudraCT. The practice of group configurations and statistical analysis could have changed over time, so that differences in the results between recently published studies and recently approved studies are not due to publication bias but due to different times.

A further limitation of the literature review was, that there was no separate analysis for phase I/II, phase II and phase II/II studies, except by the analysis of the sample size. Group configurations and statistical designs may differ between phase I/II, phase II and phase II/II. But because of the small number of phase I/II and phase II/II, I decided myself against a separate analysis.

6.Conclusion

In conclusion, although there are many theoretical proposed designs for different drug candidates, endpoints, and aims, in practice, only a few of them are used in practice. Reasons for this are, that many designs are complex, difficult to understand and implement, and organizational difficulties, e.g., in interim analysis. Furthermore, studies with a commonly used design are more likely of being approved than rarely used, complicated designs. Besides the multitude of available theoretical statistical designs, there are many different endpoints and many possibilities of group configurations. In theory, there are precise recommendations for its use depending on the advantages and disadvantages. In practice, this theoretical recommended use of the endpoints and group configurations cannot be seen, for example is ORR the most common endpoint for the evaluation of targeted and immunotherapeutic drugs, although ORR is not recommended for this use. Compared with literature reviews of 2000 and 2014, some improvement in the design and analysis and its reporting in the statistical part of phase II studies can be seen, but there is much space for further improvement. To my knowledge, there are no official guidelines for designing and analyze phase II studies, but only numerous publications dealing with the advantages and disadvantages of single aspects of group configuration and statistical designs of phase II studies. The ICH E9 guidelines aim on statistical principles of clinical trials in general, but explicit guidelines for phase II clinical studies are lacking. Better education and guidelines for designing and analyzing phase II studies are recommended, which may improve the choice of appropriate group configurations and endpoints and quality of the statistical analysis.

7. References

1. A'Hern, R.P. Sample size tables for exact single-stage phase II designs. *Stat Med* **20**, 859-866 (2001).
2. Administration), F.F.a.D. Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics Guidance for Industry. (ed. Services, U.S.D.o.H.a.H.) (2018).
3. Ananthakrishnan, R. & Menon, S. Design of oncology clinical trials: a review. *Crit Rev Oncol Hematol* **88**, 144-153 (2013).
4. Ang, M.K., Tan, S.B. & Lim, W.T. Phase II clinical trials in oncology: are we hitting the target? *Expert Rev Anticancer Ther* **10**, 427-438 (2010).
5. Brown, S.R., Gregory, W.M., Twelves, C.J. & Brown, J.M. A Practical Guide to Designing Phase II Trials in Oncology. in *A Practical Guide to Designing Phase II Trials in Oncology*, Vol. 1 (Wiley, 2014).
6. Bryant, J. & Day, R. Incorporating Toxicity Considerations Into the Design of Two-Stage Phase II Clinical Trials. *Biometrics* **51**, 1372-1383 (1995).
7. Cannistra, S.A. Phase II Trials in Journal of Clinical Oncology. *Journal of Clinical Oncology* **27**, 3073-3076 (2009).
8. Case, L.D. & Morgan, T.M. Design of Phase II cancer trials evaluating survival probabilities. *BMC Medical Research Methodology* **3**, 6 (2003).
9. Chow, S.C. & Chang, M. Adaptive design methods in clinical trials - a review. *Orphanet J Rare Dis* **3**, 11 (2008).
10. Conaway, M.R. & Petroni, G.R. Bivariate Sequential Designs for Phase II Trials. *Biometrics* **51**, 656-664 (1995).
11. Delgado, A. & Guddati, A.K. Clinical endpoints in oncology - a primer. *Am J Cancer Res* **11**, 1121-1131 (2021).
12. Dent, S.*, et al.* Application of a New Multinomial Phase II Stopping Rule Using Response and Early Progression. *Journal of Clinical Oncology* **19**, 785-791 (2001).
13. Dhani, N., Tu, D., Sargent, D.J., Seymour, L. & Moore, M.J. Alternate Endpoints for Screening Phase II Studies. *Clinical Cancer Research* **15**, 1873-1882 (2009).
14. Dignam, J., Karrison, T.G. & Bryant, J. Design and Analysis of Oncology Clinical Trials. in *Oncology. An Evidence-Based Approach* (eds. Chang, A.E.*, et al.*) (Springer, New York, NY, 2006).
15. Eisenhauer, E.A.*, et al.* New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer* **45**, 228-247 (2009).
16. Ellenberg, S.S. & Eisenberger, M.A. An efficient design for phase III studies of combination chemotherapies. *Cancer Treat Rep* **69**, 1147-1154 (1985).
17. Ellimoottil, C., Vijan, S. & Flanigan, R.C. A primer on clinical trial design. *Urologic Oncology: Seminars and Original Investigations* **33**, 116-121 (2015).
18. Ensign, L.G., Gehan, E.A., Kamen, D.S. & Thall, P.F. An optimal three-stage design for phase II clinical trials. *Statistics in Medicine* **13**, 1727-1736 (1994).
19. Farley, J. & Rose, P.G. Trial design for evaluation of novel targeted therapies. *Gynecologic Oncology* **116**, 173-176 (2010).
20. Fleming, T.R. One-sample multiple testing procedure for phase II clinical trials. *Biometrics* **38**, 143-151 (1982).
21. Gehan, E.A. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *Journal of Chronic Diseases* **13**, 346-353 (1961).
22. Grayling, M.J., Dimairo, M., Mander, A.P. & Jaki, T.F. A Review of Perspectives on the Use of Randomization in Phase II Oncology Trials. *JNCI: Journal of the National Cancer Institute* **111**, 1255-1262 (2019).
23. Guan, S. Statistical designs for early phases of cancer clinical trials. *J Biopharm Stat* **22**, 1109-1126 (2012).

24. Heitjan, D.F. Bayesian interim analysis of phase II cancer clinical trials. *Stat Med* **16**, 1791-1802 (1997).
25. Herson, J. & Carter, S.K. Calibrated phase II clinical trials in oncology. *Stat Med* **5**, 441-447 (1986).
26. Hess, K.R. Statistical issues in clinical trial design. *Curr Oncol Rep* **9**, 55-59 (2007).
27. Huang, X., Biswas, S., Oki, Y., Issa, J.P. & Berry, D.A. A parallel phase I/II clinical trial design for combination therapies. *Biometrics* **63**, 429-436 (2007).
28. Inoue, L.Y.T., Thall, P.F. & Berry, D.A. Seamlessly Expanding a Randomized Phase II Trial to Phase III. *Biometrics* **58**, 823-831 (2002).
29. Ivanova, A*., et al.* Nine-year change in statistical design, profile, and success rates of Phase II oncology trials. *J Biopharm Stat* **26**, 141-149 (2016).
30. Kilickap, S*., et al.* Endpoints in oncology clinical trials. *J buon* **23**, 1-6 (2018).
31. Kramar, A., Potvin, D. & Hill, C. Multistage designs for phase II clinical trials: statistical issues in cancer research. *Br J Cancer* **74**, 1317-1320 (1996).
32. Lee, J.J. & Chu, C.T. Bayesian clinical trials in action. *Stat Med* **31**, 2955-2972 (2012).
33. Lee, J.J. & Feng, L. Randomized Phase II Designs in Cancer Clinical Trials: Current Status and Future Directions. *Journal of Clinical Oncology* **23**, 4450-4457 (2005).
34. Lin, Y. & Shih, W.J. Adaptive two-stage designs for single-arm phase IIA cancer clinical trials. *Biometrics* **60**, 482-490 (2004).
35. Liu, P.J., Moon, J. & Crowley, J.J. Phase II selection designs. in *Handbook of Statistics in Clinical Oncology* (eds. John, C. & Antje, H.) (CRC Press, 2012).
36. López, M.F., Dupuy, J.-F. & Gonzalez, C.V. Effectiveness of adaptive designs for phase II cancer trials. *Contemporary Clinical Trials* **33**, 223-227 (2012).
37. Pallmann, P*., et al.* Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Med* **16**, 29 (2018).
38. Perrone, F*., et al.* Statistical design in phase II clinical trials and its application in breast cancer. *The Lancet Oncology* **4**, 305-311 (2003).
39. Ratain, M.J. & Sargent, D.J. Optimising the design of phase II oncology trials: The importance of randomisation. *European Journal of Cancer* **45**, 275-280 (2009).
40. Rosner, G.L., Stadler, W. & Ratain, M.J. Randomized Discontinuation Design: Application to Cytostatic Antineoplastic Agents. *Journal of Clinical Oncology* **20**, 4478-4484 (2002).
41. Rubinstein, L. Phase II design: history and evolution. *Chin Clin Oncol* **3**, 48 (2014).
42. Rubinstein, L., Crowley, J., Ivy, P., Leblanc, M. & Sargent, D. Randomized phase II designs. *Clin Cancer Res* **15**, 1883-1890 (2009).
43. Rubinstein, L.V., Gail, M.H. & Santner, T.J. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Diseases* **34**, 469-479 (1981).
44. Rubinstein, L.V*., et al.* Design issues of randomized phase II trials and a proposal for phase II screening trials. *J Clin Oncol* **23**, 7199-7206 (2005).
45. Sargent, D.J. & Taylor, J.M. Current issues in oncology drug development, with a focus on Phase II trials. *J Biopharm Stat* **19**, 556-562 (2009).
46. Schaid, D.J., Ingle, J.N., Wieand, S. & Ahmann, D.L. A design for phase II testing of anticancer agents within a phase III clinical trial. *Controlled Clinical Trials* **9**, 107-118 (1988).
47. Schlesselman, J.J. & Reis, I.M. Phase II clinical trials in oncology: strengths and limitations of two-stage designs. *Cancer Invest* **24**, 404-412 (2006).
48. Sill, M.W., Rubinstein, L., Litwin, S. & Yothers, G. A method for utilizing co-primary efficacy outcome measures to screen regimens for activity in two-stage Phase II clinical trials. *Clin Trials* **9**, 385-395 (2012).
49. Simon, R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **10**, 1-10 (1989).

50. Simon, R., Wittes, R.E. & Ellenberg, S.S. Randomized phase II clinical trials. *Cancer Treat Rep* **69**, 1375-1381 (1985).
51. Storer, B.E. A sequential phase II/III trial for binary outcomes. *Stat Med* **9**, 229-235 (1990).
52. Sun, L.Z., Chen, C. & Patel, K. Optimal two-stage randomized multinomial designs for Phase II oncology trials. *J Biopharm Stat* **19**, 485-493 (2009).
53. Tan, S.B. & Machin, D. Bayesian two-stage designs for phase II clinical trials. *Stat Med* **21**, 1991-2012 (2002).
54. Thall, P.F. & Simon, R. Incorporating historical control data in planning phase II clinical trials. *Statistics in Medicine* **9**, 215-228 (1990).
55. Thall, P.F. & Simon, R. A Bayesian approach to establishing sample size and monitoring criteria for phase II clinical trials. *Control Clin Trials* **15**, 463-481 (1994).
56. Thezenas, S., Duffour, J., Culine, S. & Kramar, A. Five-year change in statistical designs of phase II trials published in leading cancer journals. *Eur J Cancer* **40**, 1244-1249 (2004).
57. Thomas, D.W.*, et al.* Clinical Developement Success Rates 2006 - 2015. (2016).
58. Thomas, D.W.*, et al.* Clinical Developement Success Rates 2006 - 2015. (2016).
59. Wages, N.A. & Conaway, M.R. Phase I/II adaptive design for drug combination oncology trials. *Stat Med* **33**, 1990-2003 (2014).
60. Wang, M.*, et al.* Integrated phase II/III clinical trials in oncology: a case study. *Clin Trials* **9**, 741-747 (2012).
61. Winter, K. & Pugh, S.L. An investigator's introduction to statistical considerations in clinical trials. *Urol Oncol* **37**, 305-312 (2019).
62. Wu, W., Shi, Q. & Sargent, D.J. Statistical Considerations for the Next Generation of Clinical Trials. *Seminars in Oncology* **38**, 598-604 (2011).
63. Yuan, Y. & Yin, G. BAYESIAN PHASE I/II ADAPTIVELY RANDOMIZED ONCOLOGY TRIALS WITH COMBINED DRUGS. *Ann Appl Stat* **5**, 924-942 (2011).

8. List of Figures

9. List of Tables

10. Abbreviation

CR complete response
DC disease control
DF disease-free survival
DP disease progression
EFS event-free survival
EN expected sample size
LAN The Lancet
LO The Lancet Oncology
JCO Journal of Clinical Oncology
MTD maximum tolerated dose
NEJM New England Journal of Medicine
OS overall survival
pCR pathological response rate
PD progressive disease
PET probability for early termination
PFS progression-free survival
PR partial response
RECIST Response Evaluation in Solid Tumors
RFS relapse-free survival
RR response rate
SD stable disease
TTF time to treatment failure
TTP time to progression
QoL quality of life
WHO…… World Health Organization

11. Appendix

Table 11: Tumor types, for which drug candidates were tested in recently published and recently approved studies

Tumor type	Recently published studies	Recently approved studies
Adrenocortical	2% (1)	0%
Bone	5 % (3)	2% (1)
Breast	15% (9)	12% (5)
Brain	5% (3)	0%
Colon/rectal	3 % (2)	5% (2)
Fibrous	2% (1)	0%
Hepato/liver	0%	2% (1)
Immune system	8% (5)	9% (4)
Kidney	10% (6)	2% (1)
Leukemia	3% (2)	5% (2)
Lung	5% (3)	26% (11)
Mantel-cell	2% (1)	2% (1)
Melanoma/skin	2% (1)	0%
Mesothelioma	2% (1)	0%
Muscle	3% (2)	0%
Ovarian	7% (4)	2% (1)
Pancreas	3% (2)	0%
Pituitary	0%	2% (1)
Prostate	5% (3)	2% (1)
Sarcoma	2% (1)	0%
Solid tumor	3% (2)	16% (7)
Stomach/Esophagus	8% (5)	2% (1)
Thyroid	2% (1)	0%
Urothelial/bladder	3% (2)	9% (4)

Declaration on Oath

I declare that I have written the bachelor thesis independently and without outside help, that I have not used any sources other than those given and have identified the passages taken from the sources used as such. This term paper has not been presented in any other course in this or any similar form.

Munich, 11.07.2021