

Open Research Online

The Open University's repository of research publications and other research outputs

Preclinical *In Vivo* Antitumor Activity Experiments: Methodological Pitfalls and a New Framework for their Design and Analysis

Thesis

How to cite:

Porcu, Luca (2020). Preclinical In Vivo Antitumor Activity Experiments: Methodological Pitfalls and a New Framework for their Design and Analysis. PhD thesis The Open University.

For guidance on citations see [FAQs](#).

© 2019 The Author

Version: Version of Record

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online's data [policy](#) on reuse of materials please consult the policies page.

oro.open.ac.uk

**PRECLINICAL *IN VIVO* ANTITUMOR ACTIVITY EXPERIMENTS:
METHODOLOGICAL PITFALLS AND A NEW FRAMEWORK FOR
THEIR DESIGN AND ANALYSIS**

Thesis submitted for the degree of Doctor of Philosophy
at the Open University, UK

Discipline of Life Sciences

by

Luca Porcu

Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan Italy

September, 2019

Abstract

Aims: Poorly designed, analyzed and reported preclinical *in vivo* experiments (*inVivoExp*) raise ethical as well as scientific concerns. It could be hypothesized that the recurring failure of apparently promising interventions to improve outcome in clinical trials has been partially caused by poor quality of statistical design and analysis (QoStat) of *inVivoExp*. This project aimed to assess and correlate QoStat with clinical activity, and to improve the statistical framework used in *inVivoExp*.

Methods: A systematic search of Medline and EMBASE databases was carried out to identify epithelial ovarian cancer clinical trials assessing the antitumor activity of candidate compounds (CC) as monotherapy. For each eligible CC, a systematic search was carried out to identify scientific papers reporting *inVivoExp* on rats and mice, in which the CC was administered as monotherapy. An ad hoc checklist was used to assess QoStat of *inVivoExp*. QoStat was correlated to the clinical activity.

Results: Fifty-two eligible CCs and 121 *inVivoExp* were identified. In 45 out of 120 (37.5%) *inVivoExp* the method of treatment assignment was not specified. The randomization type was specified in 3 out of 74 (4.1%) *inVivoExp* and sample size was justified in 9 (7.4%) *inVivoExp*. If the primary outcome was tumor volume, the antitumor activity endpoint was declared in 14 out of 106 (13.2%) *inVivoExp*. The length of follow-up was specified in 43 (35.5%) *inVivoExp*. Outcome assessor was blinded in 5 (4.1%) *inVivoExp*. Inefficient statistical methods were often applied to analyze tumor growth data. A new statistical framework based on the Mann-Whitney statistic was proposed and applied to a specific tumor model.

Conclusions: QoStat of *inVivoExp* was so poor that the correlation with clinical activity was impossible. The magnitude of the biological signal was poorly estimated. The new statistical framework should be considered for the design and analysis of *in vivo* tumor growth studies.

“Ringrazio gli uomini di essere così buoni,
per aver dato tante manifestazioni d’amore, non riconosciuto”
Pier Paolo Pasolini

Acknowledgements

In the following rows I give thanks to people who, without them, my project wouldn't be realized. Their contributions have been different in quality and quantity, but the following values were the same to all of them: kindness, openness and modesty. No real education could be given and be received without these values.

Especially, many thanks to Dr. Roberta Frapolli. Her contribution to define the statistical design of the project was decisive. In addition, she taught me primary characteristics of *in vivo* experiments and animal models used to evaluate candidate compounds in Oncology. Moreover, whenever I had doubts about published *in vivo* experiments, she saved me, solving these doubts clearly and satisfactorily. Hundreds of times I asked her for guidance. I don't remember one single occasion that she was not available or didn't resolve my doubts.

Then, I thank Professor Maurizio D'Incalci and Professor Silvio Garattini. I would have never taken this prestigious PhD course without their support and encouragement. They are important, brilliant and well-known researchers, and yet they have been always available to speak to me and discuss my problems with kindness, openness and modesty.

I am very grateful to my supervisor, Dr. Valter Torri. My life would have been totally different without meeting him. In a very difficult moment of my life due to personal health problems, he welcomed me in his laboratory. Then he gave me the possibility and encouraged me to study science. This unique opportunity given to me by Dr. Torri, not only made me a scientist, but above all, was surely decisive in winning the battle against my health problems.

I wish to thank my examiners Dr. Nicholas Galwey and Dr. Ettore Beghi, as their support has been crucial in correcting relevant mistakes.

Special thanks to Dr. Mauro Cortellini, Dr. Alice Casagrande and Dr. Daniela Albertini. They contributed in retrieving data from biomedical literature and helped me to organize all the activities. Mauro helped me to manage my time giving me deadlines to meet during the project. I remember the first time I met Daniela. She had to prepare a thesis for a bachelor's degree in biology. She was confused when she started to read publications of *in vivo* experiments. But after only one month, she became the master and I was the scholar!

I am very grateful to my nephew Marvin Tchangwa. He is 17 years old and lives in England. He corrected the drafts of my PhD thesis discovering some mistakes in my grammar. I suppose that he was bored in doing this job. And yet, he was very kind to help me without getting bothered of me. God bless him!

Finally, I am blessed with the comfort my mother and little dog Charlie gave me during this challenging period; I could not have survived without them! Their love helped me to overcome difficulties. In the first two years of the PhD project I was worried because it seemed impossible to achieve, whilst in the last year of the PhD project I had to heavily increase my dedication towards it and consequently my mood was often cloudy. In all these situations, their presence assured me.

Contents

Abstract	2
Acknowledgements.....	4
Glossary of symbols and abbreviations	9
Chapter 1	12
Introduction.....	12
1.1 Drug development in Oncology.....	12
1.2 Principles of methodology of preclinical <i>in vivo</i> experiments.....	17
1.2.1 Internal validity.....	21
1.2.2 Reproducibility.....	22
1.2.3 Control of biological and experimental variability	24
1.2.4 External validity	25
1.2.5 R as reduction	25
1.2.6 Publication bias.....	26
1.3 Statistical analysis of <i>in vivo</i> tumor growth curves	27
1.3.1 Comparison of tumor growth curves at a selected time point	29
1.3.2 Summary statistics.....	30
1.3.3 Time-to-event endpoints.....	31
1.3.4 Multivariate methods.....	32
Chapter 2	34
Aims	34
Chapter 3	37
Methods	37
3.1 Survey design.....	37
3.1.1 Stage 1: identification of eligible CCs and estimation of their clinical antitumor activity ..	37
3.1.2 Stage 2: systematic review of <i>in vivo</i> experiments and their methodological evaluation..	40
3.2 Quality of statistical design and analysis checklist	41
3.2.1 Repetition and external validity	42
3.2.2 Internal validity.....	42
3.2.3 Statistical design	42
3.2.4 Sample size	44
3.2.5 Outcomes and their assessment	45
3.2.6 Statistical analysis.....	45

3.2.7	Attrition bias about tumor growth curves.....	45
3.2.8	Miscellanea.....	46
3.3	Sample size	46
3.4	Statistical analysis.....	47
Chapter 4	49
Results	49
4.1	Selection of clinical trials and CCs	49
4.2	Patient characteristics and assessment of clinical antitumor activity.....	53
4.3	Selection of preclinical <i>in vivo</i> antitumor activity studies	61
4.4	Tumor models used in preclinical <i>in vivo</i> antitumor activity studies	65
4.5	Assessment of the quality of statistical design and analysis.....	68
4.6	Correlation between preclinical quality and phase 2 activity	81
Chapter 5	82
Improving statistics of <i>in vivo</i> tumor growth curves		82
5.1	Concepts	82
5.1.1	Testing statistical hypotheses.....	82
5.1.2	Point and interval estimates.....	84
5.2	Non-parametric Two-Sample Tests	85
5.2.1	Definition of statistical tests	85
5.2.2	Censoring and missing data.....	87
5.2.3	Weighted non-parametric Two-Sample Tests.....	88
5.2.4	Stratified non-parametric Two-Sample Tests.....	90
5.2.5	Paired data.....	91
5.3	Statistical power and sample size determination	92
5.3.1	Location shift model	92
5.3.2	Asymptotic power	97
5.3.3	Other approximations and exact distribution of test statistic	101
5.4	An example of statistical analysis of tumor growth curves.....	101
5.5	Estimating the treatment effect.....	104
5.5.1	Introduction	104
5.5.2	Definition and properties of the estimator Δ_r^{med}	104
5.5.3	Definition and properties of the estimator Δ_r^{mean}	105
5.5.4	Summary estimators and heterogeneity between time intervals	108
5.5.5	Estimating relative effects	108
Chapter 6	109
Discussion	109

References	115
Appendix A.....	124
A.1 Preclinical search string used in the Medline database.....	124
A.2 Preclinical search string used in the EMBASE database.....	128
Appendix B.....	135
Appendix C.....	139
List of eligible clinical trials.....	139
Appendix D.....	149
List of eligible preclinical <i>in vivo</i> experiments.....	149
Appendix E	158
SAS MACRO programs	158
Appendix F	179
Tumor volumes (mm ³) measured during the <i>in vivo</i> experiment with ML017/ET myxoid liposarcoma PDX	179

Glossary of symbols and abbreviations

$1 - \beta$	Power of a hypothesis test
α	Type I error
β	Type II error
Δ_r	Additive treatment effect at the time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$
Δ_r^{mean}	Estimator of the parameter Δ_r , $r=0, \dots, K-1$
Δ_r^{med}	Estimator of the parameter Δ_r , $r=0, \dots, K-1$
$\underline{\Delta}$	Vector of additive treatment effects at time intervals $[t_r, t_{r+1}]$, $r=0, \dots, K-1$
Φ	Standard Normal Cumulative Distribution Function
aAUC	Adjusted area under the curve
AACR	American Association for Cancer Research
ASCO	American Society of Clinical Oncology
ANOVA	Analysis of variance
ARRIVE	Animal Research: Reporting <i>In Vivo</i> Experiments guidelines
AUC	Area under the curve
C	Control arm
CA-125	Cancer antigen 125
CC	Candidate compound
CDX	Cell line derived tumor xenograft
CI	Confidence interval
CR	Complete Response
$C_r(x)$	Continuous probability distribution function of the control (C) arm, for all $x \in (-\infty, +\infty)$, at the time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$
DCR	Disease Control Rate
DNA	Deoxyribonucleic acid
e&a	eligible and assessed
ECOG	Eastern Cooperative Oncology Group
EM	Expectation maximization algorithm
EOC	Epithelial ovarian cancer
EORTC	European Organization for Research and Treatment of Cancer
E_r	Expected value of the Mann-Whitney statistic, at the time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$

EU	European Union
FDA	Food and Drug Administration
GCIG	Gynecologic cancer intergroup
GOG	Gynecologic Oncology Group
HR	Hazard Ratio
<i>InVivoExp</i>	preclinical <i>in vivo</i> experiments
IQR	Interquartile range
LCK	Log cell kill
MACRO	SAS MACRO program
MASS	Morphology, attenuation, aize, and structure criteria
M-H	Mantel-Haenzel test
NCI	USA National Cancer Institute
OR	Odds Ratio
ORR	Objective Response Rate
OS	Overall Survival
PD	Progression Disease
PDX	Patient derived tumor xenograft
PFS	Progression-free Survival
PR	Partial Response
PRISMA	Preferred reporting items for systematic reviews and meta-analyses
PS	Performance Status
PSA	Prostate-Specific Antigen
QoStat	Quality of statistical design and analysis
RCB	Randomized Complete Block design
RCT	Randomized and controlled clinical trial
RECIST	Response Evaluation Criteria In Solid Tumors
RevMan	Review Manager
RoB	Risk of bias tool
SAS	Statistical Analysis Software
SD	Stable Disease
SE	Standard error
STD	Standard deviation

SWOG	South West Oncology Group
SYRCLE	Systematic review center for laboratory animal experimentation
T	Treatment arm
TGD	Tumor growth delay
TTP	Time to progression
$T_r(x)$	Continuous probability distribution function of the treatment (T) arm, for all $x \in (-\infty, +\infty)$, at the time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$
T/C	T/C ratio, where T and C are the means, or medians, of the tumor volumes of the treatment (T) and control (C) arms
UK	United Kingdom
USA	United States of America
WHO	World Health Organization
W_r	Mann-Whitney statistic at the time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$

Chapter 1

Introduction

1.1 Drug development in Oncology

Drug development is the process of bringing a new pharmaceutical drug to market once a candidate compound (CC, i.e. new chemical entity) has been identified. This process is essentially a set of applied methodologies that cover a wide range of objectives: the identification of targets, the identification of drug concentrations required for targets' inhibition and modulation, the assessment of drug pharmacokinetics and pharmacodynamics, the assessment of safety, activity and favorable or negative effects on clinical endpoints. This process is an interdisciplinary endeavour involving a multitude of professional figures from biologists, chemists, computer scientists, medical staff, statisticians, and regulatory experts. This process is time consuming and expensive. It can take 10 to 15 years and an average estimated cost exceeding USA \$1 billion (Morgan *et al.*, 2011; Rick, 2015; DiMasi *et al.*, 2016). This process is also competitive. The purpose of drug development is to select from millions of CCs those that most effectively and safely offer clinical benefit. Finally, this process is made up of a preclinical testing phase, in which *in silico*, *in vitro* and *in vivo* models are used, and a human testing phase, in which studies are conducted on human beings (i.e. clinical trials).

What advantages are there to use preclinical models? First of all, simplifications and controllability are obtained. Hence, a mechanistic insight into the impact of CC on the evolution of a disease could be obtained. Second, biological science provides explicit justification to study diseases abstracted from the entire human organism. For example, it is well known that the essential elements of tumor growth lie within cells. Cancer cells have defects in regulatory circuits that govern normal cell proliferation and homeostasis (Hanahan *et al.*, 2000). Hence, isolated cells or cell cultures are suitable objects for cancer research. As a third and last point, ethical and economic considerations request the use of preclinical models. Preliminary information about CCs' safety and efficacy profile must be collected before a CC could be reasonably administered to a human being. Without this preliminary information obtained from preclinical models, it would be unethical to test unproven chemicals in humans (Garattini *et al.*, 2017). Of course, these models must not be separated so far from reality that relevance to the ultimate goals of being better able to prevent the disease or improve treatment is lost, remembering that relevance of a result may not be evident initially.

In Oncology, the classic approach taken to identify chemotherapy drugs, requires that the CC is first evaluated against a panel of malignant cell lines, such as those used by USA National Cancer Institute (NCI-60, refer to the web site: https://dtp.cancer.gov/discovery_development/nci-60/cell_list.htm. Developmental Therapeutics Program. National Cancer Institute. Last Updated: 8 May 2015. Retrieved 01 September 2019). If the CC shows antitumor activity in the panel of cell lines, other *in vitro* studies are performed to determine its mechanism of action. New methods develop CCs in a different manner, namely targeting specific molecules or pathways known to have a role in tumor growth. These CCs can be identified in different ways, such as the screening of small-molecule libraries or by computer-assisted protein-structured-based design. A biochemical or cellular assay is then required to evaluate the effects on the molecular target and those CCs that should undergo further development are selected.

Biological differences between primary tumors and the cancer cell lines derived from them, limit the value of *in vitro* studies for the evaluation of CCs (Szakács *et al.*, 2004). Once the target and mechanism of action have been identified using *in vitro* models, *in vivo* experiments are undertaken to ensure that inhibition of the target can be achieved at tolerated doses *in vivo* and to identify and validate predictive biomarkers of response. Chemotherapy drugs can be considered ‘targeted’ in that they inhibit DNA synthesis and the cell division apparatus. The theory behind the preclinical *in vivo* experimentation is to look for activity in *in vivo* models which would translate into some likelihood of activity in human disease. *In vivo* models are also required to evaluate CCs’ pharmacokinetics, CCs’ effects on biological processes such as invasion into neighboring tissues, angiogenesis, metastasis, and the relative effects of CCs against tumor cells compared to their toxicity in normal tissues (Ocana *et al.*, 2010).

Once preclinical *in vivo* experiments have been successfully performed, CCs could be administered in humans for the first time. When CCs reach the clinical setting, drug development proceeds through a series of sequential clinical phases designed to assess their safety and efficacy. The standard clinical paradigm for the evaluation of CCs consists of a phase I trial to establish the optimal dose, a phase II trial to obtain preliminary evidence of activity and a phase III trial for comparison with the standard therapy. In this approach, the phases I and II (exploratory trials) are for gathering information and screening the CC; the phase III (confirmatory trial) is for a definitive comparison with the standard therapy.

In Oncology, phase II trials are an essential bridge between the small phase I trials, which determine the dose of antitumor CCs, and the large-scale and confirmatory phase III trials. The

primary aim of phase II trials is to screen CCs for their biologic antitumor activity. Secondary aims are the preliminary evaluation of CCs' safety profile and predictive biomarkers.

CCs' antitumor activity is assessed using standardized response criteria. In solid tumors, the first international standardized response criteria were written and disseminated by the World Health Organization (WHO) in 1979 [World Health Organization. WHO Handbook for Reporting Results of Cancer Treatment Offset Publication No. 48. (WHO press, Geneva, 1979)]. Of primary relevance, the authors defined exactly what constitutes a response to treatment or progression of disease by standardizing the amount of tumor shrinkage necessary to qualify a patient for each of four categories: complete response (CR: disappearance of all known disease by two observations at least 4 weeks apart), partial response (PR: 50% or more decrease in total tumor size of the lesions that have been measured. No new lesions. No progression of any lesion), stable disease (SD: it cannot be established that the total size has decreased by at least 50%, nor has a 25% increase in the size of one or more measurable lesions been demonstrated) and progressive disease (PD: a 25% or more increase in the size of one or more measurable lesions or the appearance of new lesions). This answered an urgent need of medical oncology. Because single-arm, uncontrolled, Phase II trials were used to assess CCs' antitumor activity, standards to compare responses across trials were urgently needed. Widespread application of the WHO criteria, however, brought to light some deficiencies/discrepancies. The reliability of the methodology both in terms of intraobserver as well as in terms of interobserver variability was questioned (Warr *et al.*, 1984; Tonkin *et al.*, 1985; Warr *et al.*, 1985; Thiesse *et al.*, 1997). Cooperative groups and pharmaceutical companies often 'modified' original WHO criteria to accommodate new technologies for human cancer imaging or to address areas that were unclear in the original document. For example, the South West Oncology Group (SWOG) published their version of the WHO criteria in 1992 (Green *et al.*, 1992). As a major change, a larger increase in tumor size (50%) was requested to define PD. In the same year, the European Organization for Research and Treatment of Cancer (EORTC) published its own version of the WHO criteria (Tumor eligibility and response criteria for phase II and III studies. Brussels: EORTC Data Center Manuel 1992), defining minimum sizes for lesions from different organs to be considered as measurable. Because different versions of the original WHO criteria were used in clinical trials, the comparison of results of clinical trials became very unreliable. In 1994, several clinical research organizations began updating the WHO standards and, 6 years later, published a new version, under the acronym RECIST (Response Evaluation Criteria In Solid Tumors, Therasse *et al.*, 2000). The WHO and RECIST standards share the same principles, that is standardizing the

amount of tumor shrinkage necessary to qualify a patient for each of the previous four classifications. In RECIST criteria, the assessment of tumor lesions was simplified and better specified in order to address apparent deficiencies and lack of details of the WHO criteria. RECIST requires that certain lesions are identified as the key lesions that will track disease change; RECIST alters the definition of PR and PD and changes the way that lesions are measured (unidimensional versus bidimensional in the WHO criteria); the RECIST standards also update which imaging modalities are acceptable for measuring tumor size. In 2009, RECIST criteria were updated in the version 1.1 (Eisenhauer *et al.*, 2009). Major changes were the reduction of the number of lesions to be assessed, how to assess pathological lymph nodes was specified, the definition of PD was better specified, the confirmation of response was not requested anymore in randomized trials and, finally, what constitutes 'unequivocal progression' of non-measurable/non-target disease was explained.

Whereas the RECIST criteria address many shortcomings of previous attempts to classify tumor response, they have limited utility in the evaluation of ovarian cancer. In recurrent ovarian cancer, a significant proportion of patients have only micro-nodular peritoneal carcinomatosis and ascites, which are non-measurable according to RECIST criteria. Because the RECIST criteria define tumor response on the basis of evaluation of measurable disease, it precludes its use in almost 50% of ovarian cancer patients (Rustin *et al.*, 2004). To allow the inclusion of these patients, it was proposed that the CA-125 serum tumor marker could be utilised as a tumor response criterion. CA125 is a high molecular weight glycoprotein which is raised in approximately 90% of patients with advanced epithelial ovarian cancer (Bast *et al.*, 1983). Many more patients are evaluable according to CA-125 than those assessed by computed tomography scanning used to assess standard (WHO or RECIST) response criteria (van der Burg *et al.*, 1993; Pearl *et al.*, 1994; Rustin *et al.*, 1996;). Moreover, measurement of CA-125 is less expensive and more comfortable for patients than computed tomography scanning. Characterized in 1981, the CA-125 antigen has several important roles in the routine management of ovarian cancer patients and could be used as a prognostic marker (Rustin *et al.*, 2004). In 1996, Rustin *et al.* defined criteria for evaluating 50% and 75% response according to CA-125. Based on retrospective studies, the Gynecologic Cancer Intergroup (GCIg) proposed that a definition of ovarian tumor progression based on CA-125 doubling should be used in clinical trials of first-line therapies (Vergote *et al.*, 2000). In addition to increasing the number of eligible patients for a given trial, it was suggested that utilisation of a composite definition of progression based on both RECIST and CA-125 criteria (instead of only one or the other) would increase the statistical power for tests of differences between trial arms regarding PFS (Rustin

et al., 2006). Thus, a public workshop sponsored by the US Food and Drug Administration, American Society of Clinical Oncology, and American Association for Cancer Research (FDA-ASCO-AACR) recommended CA-125 to be used as a surrogate marker of disease progression (Bast *et al.*, 2007). They also proposed that CA-125 should be included as a part of a composite endpoint that includes radiological and clinical evaluation.

There are two main types of endpoints based on standardized response criteria: binary and time-to-event. Binary endpoints include the Objective Response Rate (ORR, i.e. the proportion of patients whose tumor exhibits a PR or CR), and the Disease Control Rate (DCR, i.e. the proportion of patients whose tumor exhibits SD, PR or CR). Time-to-event endpoints include the progression-free survival (PFS) and the time-to-progression (TTP): the former measures time-to-tumor progression or death whichever occurs first, while the latter treats death as a censoring event. Based on antitumor mechanism, different endpoints could be used to detect CC's antitumor activity. Broadly, ORR is considered suitable for cytotoxic drugs but less suitable for cytostatic agents (Adjei *et al.*, 2009; Sharma *et al.*, 2012). PFS is said to be more informative for cytostatic agents (Seymour *et al.*, 2010).

A recent survey (Hay *et al.*, 2014) demonstrated that Oncology has one of the highest attrition rates in the drug development process. Oncology drugs have the lowest likelihood of success from phase I; only around 1 in 15 drugs (6.7%, $n = 1.803$) of all indication development paths in phase I were approved by FDA. In particular, the phase II success rate (i.e. the probability of a drug moving from phase II to phase III) was estimated at 28.3%. The unsatisfactory positive predictive value of phase II trials (i.e. the low probability of reaching market approval from the phase II) is explained by the following reasons:

- The strength of activity signal obtained in phase II Oncology trials is often too low to cause a clinical benefit in large-scale and confirmatory phase III trials
- The methodology applied to phase II Oncology trials is generally low-level. First, there is a lack of surrogate biomarkers that can be measured earlier than survival, and that can predict phase III outcome more reliably than conventional response criteria based on tumor size variations. Correlation with clinical endpoints does not mean surrogacy. Exactly 30 years ago, Prentice formulated the criteria to demonstrate the surrogacy of a biomarker (Prentice, 1989). These criteria require thousands of patients enrolled in different clinical trials. To date, there are only a handful of accepted biomarkers that are established surrogate endpoints. In prostate cancer, for example, the prostate specific antigen (PSA) decrease has

been reasonably well validated in Phase III studies of cytotoxic agents, although there is debate on using this biomarker in exploratory trials (Stadler, 2002; Williams, 2018). Second, poor quality statistical designs have been traditionally used in phase II Oncology trials. The traditional single-arm phase II Oncology trial uses a historical response rate as the reference point by which improved response rate is judged. Outcomes of single-arm phase II trials reflect some combination of treatment effect, random effect, and unknown differences between treated and historical control patients. Recommendations have been produced to use randomization to protect against selection bias in phase II Oncology trials (Booth *et al.*, 2008; Ratain *et al.*, 2009). Also dose-ranging, controlled phase II trials should receive considerable attention in order to determine the relationship between dose and CCs' antitumor activity (Ratain, 2005; Michaelis *et al.*, 2006). Finally, blinding techniques could be useful to prevent different types of biases (i.e. performance, assessment, and attrition biases. Table 1.2.1.1 reports their definition), especially for time-to-event endpoints such as PFS and TTP. Unfortunately, it is often difficult to apply blinding techniques in Oncology. For example, to mask devices, routes of administration and side effects such as myelosuppression or nausea is often impossible and unethical

- CCs are wrongly selected in the preclinical drug development or, at least, the positive predictive value (i.e. the probability of reaching market approval from the preclinical testing phase) of methodologies applied in the preclinical drug development, is unsatisfactory.

1.2 Principles of methodology of preclinical *in vivo* experiments

Animal experiments remain essential to understand the fundamental mechanisms underpinning malignancies and to discover and screen methods to prevent, diagnose and treat them. Given the limited usefulness and predictive capability of *in silico* and *in vitro* models, the use of animal models must continue (Garattini *et al.*, 2017). In Europe, animal research is tightly controlled under the European Directive 2010/63/EU; ethical validity is usually judged in relation to the “three Rs” (i.e. Replacement, Reduction and Refinement) introduced by Russell and Burch in their book, “The Principles of Humane Experimental Technique”, first published in 1959 (Russell *et al.*, 1992).

Michael Festing, one of the most prominent statistician involved in animal research, reminds us that “the use of animals in biomedical research generates strong emotions, but everyone will surely agree that if they are used the experiments should be properly designed and cause the minimum amount of pain and distress” (Festing, 2010). And yet, a recent survey of 271 papers from academic organisations in the UK and USA involving work on live laboratory mice, rats or non-human

primates, has found that the design, analysis and reporting of animal experiments could be improved (Kilkenny *et al.*, 2009). The survey's major findings are reported in Table 1.2.1 (Festing, 2010). That survey has spawned a follow-up paper introducing the ARRIVE guidelines (Kilkenny *et al.*, 2010). ARRIVE stands for Animal Research: Reporting *In Vivo* Experiments. ARRIVE guidelines were published first in *PLoS Biology* and then in several other journals. These guidelines consist of a checklist of 20 items describing the minimum information that all scientific publications reporting research using animals should contain.

Survey finding	Percentage of studies
Purpose of the study not clearly stated in the introduction	5
Did not clearly indicate how many separate experiments were done	6
Failed clearly to identify the experimental unit	13
Failed to state the sex of the animals	26
Reported neither age nor weight of animals	24
Failed to record the exact number of animals used (although in several cases an approximate number could be estimated)	36
Failed to justify the sample sizes used	100
Reports of the numbers of animals used differed between materials and methods and results sections	35
Random allocation of animals reported	12
Studies reporting blinding when qualitative scoring was used	14
Studies where the statistical methods used were not clear or not reported	4
Studies with numerical data which failed to present a measure of variation such as a standard deviation, standard error, or confidence interval	17
Papers judged not to have used the correct statistical methods, or where the methods used were not clear	12

Table 1.2.1 Primary findings of the survey of the quality of experimental design, statistical analysis and reporting of research using animals. 271 papers from academic organisations in the UK and USA were assessed (Festing, 2010)

The poorly designed, analysed and reported preclinical *in vivo* experiments raise ethical and scientific concerns about proper use of animals and reproducibility, respectively. Key methodological issues of preclinical *in vivo* experiments are shown in Figure 1.2.1. All these issues are put into doubt by variability of experimental results, measurements and biological models. They are summarised in the following sections.

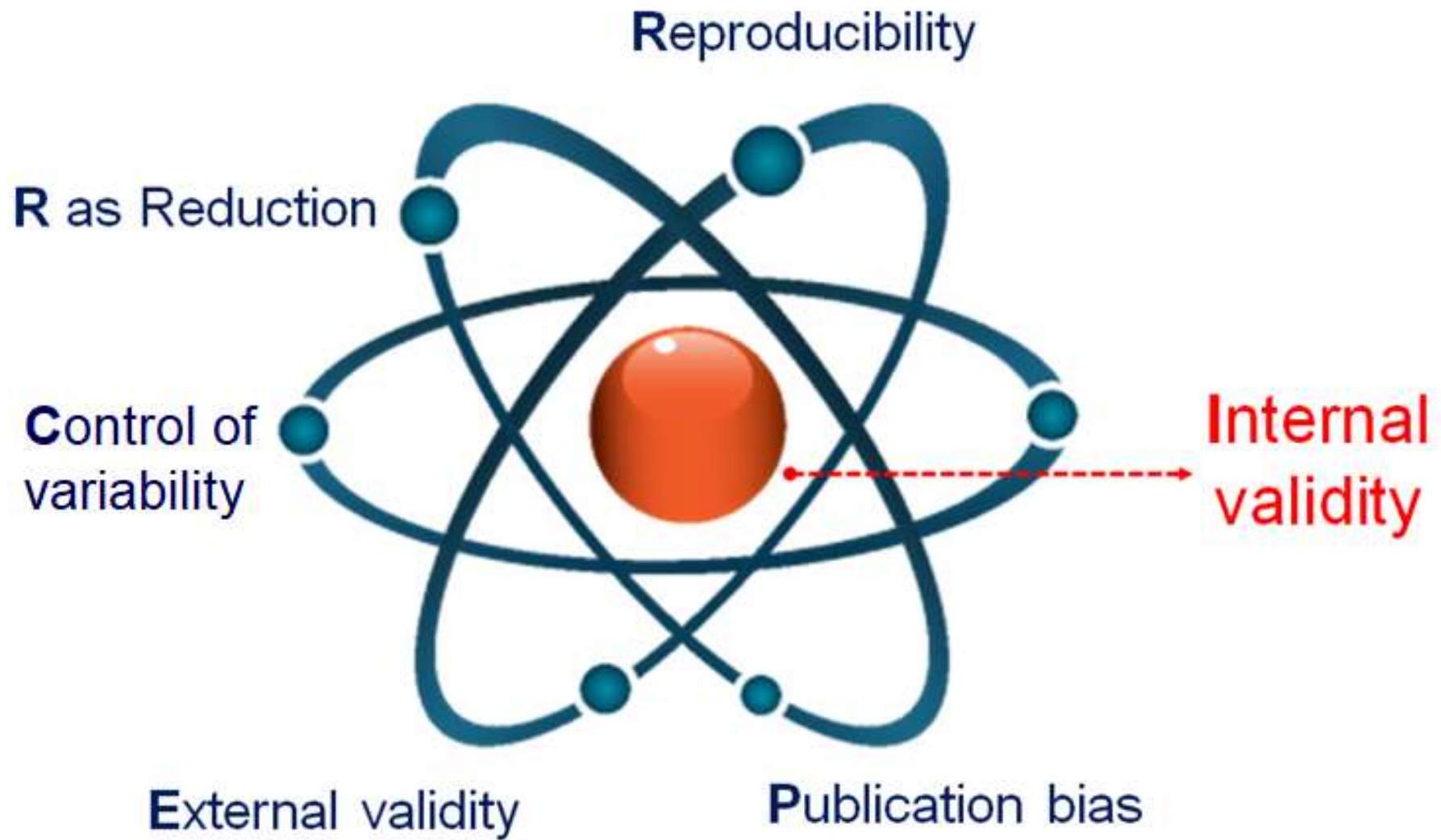


Figure 1.2.1 Methodological issues of preclinical *in vivo* experiments

1.2.1 Internal validity

Internal validity is the core issue. A preclinical *in vivo* experiment with poor internal validity implies poor reproducibility. Due to poor reproducibility, its results are suspiciously accepted by the scientific community. The situation is worse still. Systematic reviews and meta-analyses of all available evidence from preclinical *in vivo* experiments produce low weight of evidence if single *in vivo* experiments have poor internal validity. Adequate internal validity of a preclinical *in vivo* experiment means that the differences observed between groups of animals allocated to different interventions may, apart from random error, be attributed to the treatment under investigation (Jüni *et al.*, 2001). By definition, random error is totally controlled by the calculus of probability. Neither the calculus of probability nor other statistical tools can handle systematic error (bias) without unverified assumptions.

Four types of bias threaten internal validity. Their definition and possible solution are reported in Table 1.2.1.1.

Type of bias	Definition	Solution
Selection bias	Treated and control groups differ prior to treatment in ways that matter for the outcomes under study	Randomization; allocation concealment; intention-to-treat analysis
Performance bias	Systematic differences in care between the treatment groups apart from the intervention under study	Blinding
Assessment/detection bias	Systematic differences between treatment groups in the assessment of study outcomes	Blinding
Attrition bias	Systematic differences between treatment groups in the number and the way animals are lost or exit from the experiment	Blinding; intention-to-treat analysis

Table 1.2.1.1 Types of bias threatening internal validity

To prevent selection bias, treatment allocation should be based on randomization. This means that an *a priori* determined probability of enrollment in a specific treatment or control group should be assigned to each animal. This is not enough. To prevent selection bias, concealing the allocation

sequence from those assigning animals to intervention groups, until the moment of assignment, should be applied. In few words, picking animals 'at random' from their cages has the risk of conscious or subconscious manipulation, and does not represent a true and satisfactory method of randomization. To prevent performance, detection, and attrition bias, caregivers, researchers and outcome assessors should be blinded from knowing which intervention each animal received during the experiment. Blinding may not always be possible in all stages of an experiment, for example when the treatment under investigation concerns a surgical procedure or the treatment safety profile unmasks the administered treatment. However, blinding of outcome assessment is almost always possible. In a retrospective review, 290 animal studies with intervention were classified by the use of randomization and blinding (Bebarta *et al.*, 2003). The Odds Ratio (OR) of reporting a significant difference was 3.4 (95%CI: 1.7 to 6.9) for the studies in which randomization was not used compared to those in which randomization was used. The OR of reporting a significant difference was 3.2 (95%CI: 1.4 to 7.7) for the studies in which blinding was not used compared to those in which blinding was used. Finally, the OR of reporting a significant difference was 5.2 (95%CI: 2.0 to 13.5) for the studies in which both experimental techniques were not used compared to the studies in which both techniques were used. These results suggest that failure to blind and randomize may lead to bias.

Intention-to-treat analysis is the analysis of data of all animals included in the group to which they were randomly assigned, regardless of whether they completed the intervention. This statistical procedure is useful to prevent selection and attrition bias. For instance, suppose that animals dead for treatment toxicity are removed from the final analysis. It could be argued that only animals with specific characteristics are retained in the final analysis and the measure of treatment effect respect to control group is biased. Risk of bias tools (SYRCLE's RoB tool) for animal intervention studies are useful to evaluate the level of internal validity of *in vivo* experiments (Hooijmans *et al.*, 2014).

1.2.2 Reproducibility

The ability to reproduce experiments is at the heart of science. Goodman *et al.* (2016) define this term in three different ways, that are reported in Table 1.2.2.1.

Type of reproducibility	Definition
Methods	Methods reproducibility refers to the provision of enough detail about study procedures and data so the same procedures could, in theory or in actuality, be exactly repeated
Results	Results reproducibility refers to obtaining the same results from the conduct of an independent study whose procedures are as closely matched to the original experiment as possible
Inferential	Inferential reproducibility refers to the drawing of qualitatively similar conclusions from either an independent replication of a study or a reanalysis of the original study. Inferential reproducibility is not identical to results reproducibility or to methods reproducibility, because scientists might draw the same conclusions from different sets of studies and data or could draw different conclusions from the same original data, sometimes even if they agree on the analytical results

Table 1.2.2.1 Types of reproducibility

Scientists in the Haematology and Oncology department at the biotechnology firm Amgen in Thousand Oaks, California, tried to confirm published findings of ‘landmark studies’ in Oncology (Begley *et al.*, 2012). Fifty-three papers were deemed ‘landmark’ studies (i.e. something completely new, such as fresh approaches to targeting cancers or alternative clinical uses for existing therapeutics). Scientific findings were confirmed in only 6 (11%) cases! This disappointing result could be due to the following reasons: poor internal validity, lack of good reporting and transparency, and poor control of biological variability. Lack of reproducibility in other laboratories may also be caused by treatment x environment interactions. For example, animal houses may differ in the physical environment, management, or microflora in such a way as to alter the relative treatment differences. These are the reasons threatening reproducibility in science. A similar finding was reported by Prinz *et al.* (2011). The scope of the Prinz *et al.* study was to compare in-house results with published results for wet-lab experiments related to drug target identification and validation. Sixty-seven in-house projects within the oncology (47 projects, 70%) , women’s health (12 projects, 18%) and cardiovascular (8 projects, 12%) indications were used to reproduce published data. Only in 20 to 25% of the projects in-house findings were completely in line with

published data. In almost two-thirds of the projects, there were inconsistencies between in-house data and published data that either considerably prolonged the duration of the target validation process or, in most cases, resulted in termination of the projects because the evidence that was generated for the therapeutic hypothesis was insufficient to justify further investments into these projects.

1.2.3 Control of biological and experimental variability

Russell and Burtch's chapter on reduction, written in 1959, is largely concerned with the control of inter-individual biological variation through the use of inbred strains (Russell *et al.*, 1992). The control of variability delivers enormous advantages in *in vivo* experimentation, that are reported in Table 1.2.3.1.

Type of benefit	Explanation
Power	Uncontrolled biological variability leads to increased numbers of false negative results. The noise (i.e. biological variability) prevails over the biological signal
Reproducibility	Uncontrolled biological and experimental variability leads to lack of methods and results reproducibility
Reduction	Controlling biological and experimental variability, the signal (i.e. treatment effect) / noise (i.e. variability) ratio is increased and less animals are necessary to detect the same treatment effect

Table 1.2.3.1 Benefits derived from the control of biological and experimental variability

One of the methods largely suggested to control biological variability has been the use of blocks. Simple randomization requires substantial numbers of animals in order to fully randomly balance all possible confounding factors (e.g. animal strain, age, gender, weight, housing). In randomized block designs different sources of variability are distributed in a controlled manner to the individual block entities to which individual animals are assigned at random. Blocks could be useful to guarantee reproducibility. Suppose that an experiment is executed in different times or laboratories. Times or laboratories could be used as blocks. If there is good agreement between these blocks, then this gives some assurance that the experiment is reproducible. Other useful statistical designs to control biological variability include Latin square, crossover designs and repeated measure design (Festing *et al.*, 1998; Festing *et al.*, 2002).

1.2.4 External validity

External validity could be defined as the extent to which the results of a preclinical *in vivo* experiment provide a correct basis for generalisations to the human condition. Ideally, a disease model should fully reproduce the clinical condition in a system that can be used for research and drug discovery. But all preclinical models are an imperfect replication and simplified models of the clinical condition. The following reasons could explain the failed translation of *in vivo* experiments to the clinic:

- Differences between *in vivo* models and humans, testing the same treatment (e.g. pathophysiology of disease, comorbidities, age)
- Differences between the treatment administered in an *in vivo* experiment and that administered in humans e.g. (timing of the administration, dosing of the study treatment, using of co-medications)
- Differences in the outcome measures (e.g. in *in vivo* antitumor activity studies, tumor growth curves are usually used to detect CCs' treatment effect. In clinical trials time-to-progression could be used to detect CCs' treatment effect)
- Shortcomings of the clinical trial. For instance, clinical trials may have had insufficient statistical power to detect a true benefit of the treatment under study or the same treatment was administered at later time points when the window of opportunity has passed (Gladstone *et al.*, 2002; Grotta, 2002).

If the issues regarding internal validity are almost the same in all *in vivo* experiments, regardless of the disease under study, the external validity of an *in vivo* experiment will largely be determined by disease-specific factors.

1.2.5 R as reduction

The number of animals used should be reduced to the minimum consistent with achieving the objectives of the preclinical *in vivo* experiment. Reduction, of course, lies squarely in the field of statistics. Table 1.2.5.1 reports the statistical techniques available to reduce sample size in *in vivo* experiments.

Statistical technique	Explanation
Increasing the signal/noise ratio	The number of animals is smaller if a larger treatment effect is targeted and/or the biological and experimental variability is reduced

Multi-arm designs	Multiple treatments could be evaluated in the same experiment. Control arm is the same for all active arms. Interactions between treatment factors could be fairly evaluated using a factorial design
Choosing appropriate endpoints	For instance, continuous endpoints are more powerful than categorical endpoints; repeated measures instead of single measures increase the power of common statistical tests
Using indirect evidences	Historical data could be combined to <i>in vivo</i> experiment's data using, for example, bayesian techniques (Gelman <i>et al.</i> , 2004). Moreover, historical data should be used to guide statistical design
Increasing statistical errors	The number of animals could be reduced by accepting more false positive (i.e. type I) and negative (i.e. type II) errors (refer to Section 5.1.1 for their explanation). For instance a type I error of 0.05 could be substituted by a type I error of 0.10 and a type II error of 0.20 could be substituted by a type II error of 0.22. Moreover, in case of <i>in vivo</i> experiments screening CCs, two-tailed tests could be substituted by one-tailed tests
Adaptive designs	If data are analyzed at interim, decision rules such as stopping rules or sample size re-estimation could be applied. Hence, the the number of animals used in <i>in vivo</i> experiments is better justified and, stopping early the experiment, is reduced

Table 1.2.5.1 Statistical techniques to reduce sample size in *in vivo* experiments

Excluding multi-arm designs, other statistical techniques are rarely applied in *in vivo* experiments. For example, randomized block designs are scarcely used (Festing, 2014). Adaptive designs are almost never applied in preclinical *in vivo* experimentation.

1.2.6 Publication bias

Systematic review and meta-analysis are techniques developed for the analysis of data from clinical trials. They may be helpful also in preclinical research. For instance, a systematic review and meta-

analysis of all available evidence from preclinical studies should be performed before clinical trials are started.

If studies are published selectively on the basis of their results, even a meta-analysis based on a rigorous systematic review will be misleading. In a meta-analysis of 525 publications included in systematic reviews of 16 interventions tested in animal studies of acute ischaemic stroke, it was estimated that publication bias might account for around one-third of the efficacy reported in systematic reviews of animal stroke studies and that a further 214 experiments, in addition to the 1,359 identified through rigorous systematic review (non publication rate: 14%), have been conducted but not reported (Sena *et al.*, 2010).

Nonpublication of the results of animal studies is unethical because the included animals are wasted. They do not contribute to accumulating knowledge. As a consequence, ‘wrong ways’ could be taken in preclinical and clinical research:

- overstated biological effects may lead to further unnecessary *in vivo* experiments testing poorly founded hypotheses
- publication bias deprives researchers of the accurate data they need to estimate the potential of novel therapies in clinical trials.

The recognition of substantial publication bias in the clinical literature has led to the introduction of clinical trial registration systems to ensure that those summarising research findings are at least aware of all relevant clinical trials that have been performed (De Angelis *et al.*, 2004). A central register of preclinical *in vivo* experiments performed should be kept along with their respective reference publications (van der Worp *et al.*, 2010).

1.3 Statistical analysis of *in vivo* tumor growth curves

In preclinical *in vivo* experiments, antitumor activity is usually evaluated by estimating the tumor volume at different times after drug administration. In a typical experiment, rodents, usually mice or rats, are inoculated subcutaneously with tumor cells that are either isogenic, if the rodent is immunocompetent, or xenogenic (i.e. human tumor cells are inoculated), if the rodent is immunodeficient. Alternatively, tumor cells can be injected orthotopically, into the organ from which they originate. Tumors could also be induced by administration of carcinogens or genetic manipulations (Zitvogel *et al.*, 2016). Rodents that develop tumors reaching a predetermined volume are randomized into different treatment and control groups and drugs are administered. Those rodents injected with tumor cells but with no sign of tumor burden are usually sacrificed after inoculation. The volume of each tumor is measured at the start of treatment and periodically

throughout the experiment. Rodents are sacrificed either when their tumor volume reaches a maximum target volume, or when a humane endpoint (i.e. the earliest, predetermined, physiological or behavioral sign used to avoid or stop the distress, discomfort, or potential pain and suffering) is reached or at the end of follow-up (administrative censoring). The resulting dataset consists of incomplete, repeated measures of tumor volume at common time points, from the start of treatment until the time in which the last rodent has been sacrificed. An example of tumor growth curve is reported in Figure 1.3.1.

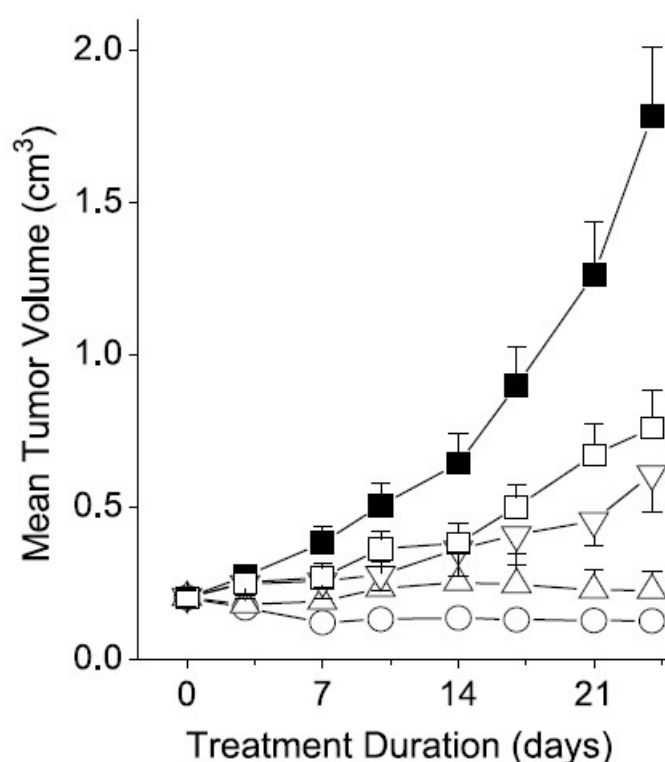


Figure 1.3.1 Antitumor effects of AZD2171 (□, 0.75 mg per kg per day; ▽, 1.5 mg per kg per day; △, 3 mg per kg per day; ○, 6 mg per kg per day) or vehicle (■) on growth of MDA-MB-231 human breast tumor xenografts. Xenografts were established s.c. in athymic mice and allowed to reach a volume of $0.2 \pm 0.01 \text{ cm}^3$ (mean \pm standard error) before treatment. Once-daily oral administration of AZD2171 or vehicle then commenced and was continued for the duration of the experiment. *Points*, mean from 10 to 11 mice; *bars*, standard error in one direction (Wedge *et al.*, 2005)

Limitations of this method include lack of information about the effects of the CCs on metastases, or the process of metastatic spread. Also, in order to evaluate the mechanism(s) of action of a drug, rodents must be killed to allow molecular analysis of the resected tumor. In addition, although tumor growth curves with and without treatment reflect tumor response or delay in progression,

these end points may not reflect selective effects against those tumor cells with high reproductive potential (e.g. putative stem cells) that are important in determining the long-term benefits of treatment (Ocana *et al.*, 2010).

Tumor volumes are measured on a weekly basis using a caliper on determined days. Imaging techniques, such as bioluminescence imaging, may be used to record changes in the volume of tumors that are not restricted to superficial sites and/or to provide information about drug-influenced biological processes (e.g. metastatic spread, expression of proteins). Details about imaging techniques and their use are reported in Ocana *et al.*, 2010.

To analyse data series of tumor volumes at different time points, the common statistical practice is first to demonstrate that the treatment influences them, then to estimate the treatment effect. To solve the former problem, statistical tests are used (Lehmann *et al.*, 2005), while to solve the latter problem, unbiased estimators are used (Lehmann *et al.*, 1998). Details about hypothesis testing and statistical estimation are reported in Section 4.3.1. The statistical approaches currently used to analyse data series of tumor volumes at different time points, could be classified in the following categories:

- Data analysis at a selected time point
- Use of summary statistics to estimate treatment effect
- Substitution of data series with the time required to reach a target volume
- Use of multivariate methods

An overview of these statistical approaches is shown below.

1.3.1 Comparison of tumor growth curves at a selected time point

Control and treatment arms are compared at a selected time point; usually the time point at the end of follow-up. The statistical test at each time point could be parametric, namely t test for two arms or ANOVA test for more than two arms, or non-parametric, namely Mann-Whitney test for two arms or Kruskal-Wallis test for more than two arms. The T/C ratio, calculated at the selected time point, is a common measure of treatment effect (Corbett *et al.*, 2003; Houghton *et al.*, 2007). T and C are the means, or medians, of the tumor volumes of the treatment (T) and control (C) arms, respectively, at the selected time point. From a statistical point of view, this approach is inadequate as explained in the following three points:

- 1) comparing control and treatment arms at a selected time point is a weaker comparison than that of the tumor growth curves over all times. It neither captures all the data nor addresses the

different biological mechanisms underlying tumor growth. The suboptimal use of data series is well represented by the following examples:

- 1a. suppose that at the end of follow-up control and treatment arms have the same tumor volume distribution but previously, tumor growth was constant in the control arm while tumor volume was greatly reduced and then quickly increased in the treatment arm, as reported in Figure 1.3.3.1. Treatment effect is not formally recognized by this statistical approach
- 1b. suppose that, at the start of treatment, there is a tumor volume reduction in the treatment arm but, after few days, tumor growth curves of control and treatment arms remain parallel to each other over all the rest of follow-up. At the end of follow-up, treatment effect could be formally recognized by this statistical approach although its biological relevance is poor
- 2) the choice of the time point could be data driven (e.g. most of rodents in the control arm are just sacrificed) or *a priori* (e.g. based on the planned treatment administration). In the first case, comparison is constrained by specific events, such as animal sacrifice in the control arm, that weaken and rather render ambiguous the interpretation of this comparison. In the second case, the *a priori* choice of the time point creates difficulty because unreliable assumptions are needed (e.g. exponential growth with a determined growth rate in the control arm) to design and determine sample size
- 3) attrition bias due to censoring animals (e.g. rodents previously sacrificed) could affect the formal comparison at the selected time point.

A worst method is to repeat this approach at different time points, indicating all the times at which differences were significant. This procedure may be naively considered better because it uses all the data series. On the contrary, due to its very bad procedure caused by the inflation of type I error due to the multiple comparisons problem, post-hoc tests are difficult to apply because repeated measures are correlated and comparisons are usually underpowered.

1.3.2 Summary statistics

Per-experiment and per-animal summary statistics are commonly used to estimate treatment effect. Examples of per-experiment summary statistics are the minimal T/C ratio, which reflects the maximal tumor growth inhibition achieved (Hendriks *et al.*, 1992), and the adjusted AUC ratio (aAUC ratio; Wu *et al.*, 2010). The minimal T/C ratio is the minimum of the T/C ratios calculated at all time points. aAUC ratio is defined as the ratio of the means of the aAUCs of the treatment and control groups, where aAUC is the per-animal area-under-the-curve (AUC) calculated up to the last time point available for the rodent, divided by the length of the interval between the start of treatment

and the last time point with existing tumor volume measurements. Another example of per-animal summary statistics is the T_{nadir} . It is defined as the minimum of the growth curve of a treated tumor relative to the tumor volume at the start of treatment (Ubezio, 2019).

Per-experiment and per-animal summary statistics are usually easy to calculate and informative. However, their sampling distribution could be highly skewed and average values such as the aAUC ratio could suffer from suboptimal power with respect to multivariate methods.

1.3.3 Time-to-event endpoints

Control and treatment arms are compared in terms of time in days for the tumors to reach a predefined target volume (tumor growth delay). For instance, it could be the doubling time of tumor volume, defined as the earliest day on which the tumor volume is at least twice as large as on the first day of treatment. The non-parametric log-rank test and the semi-parametric Cox regression model are available to detect and estimate treatment effect, respectively, in the presence of right-censored data. However, there are two major disadvantages when using this approach. First, the choice of the target volume at which to assess the delay is critical for this comparison (Begg, 1980). Second, it neither captures all of the data nor addresses the different biological mechanisms underlying tumor growth, as reported in Figure 1.3.3.1.

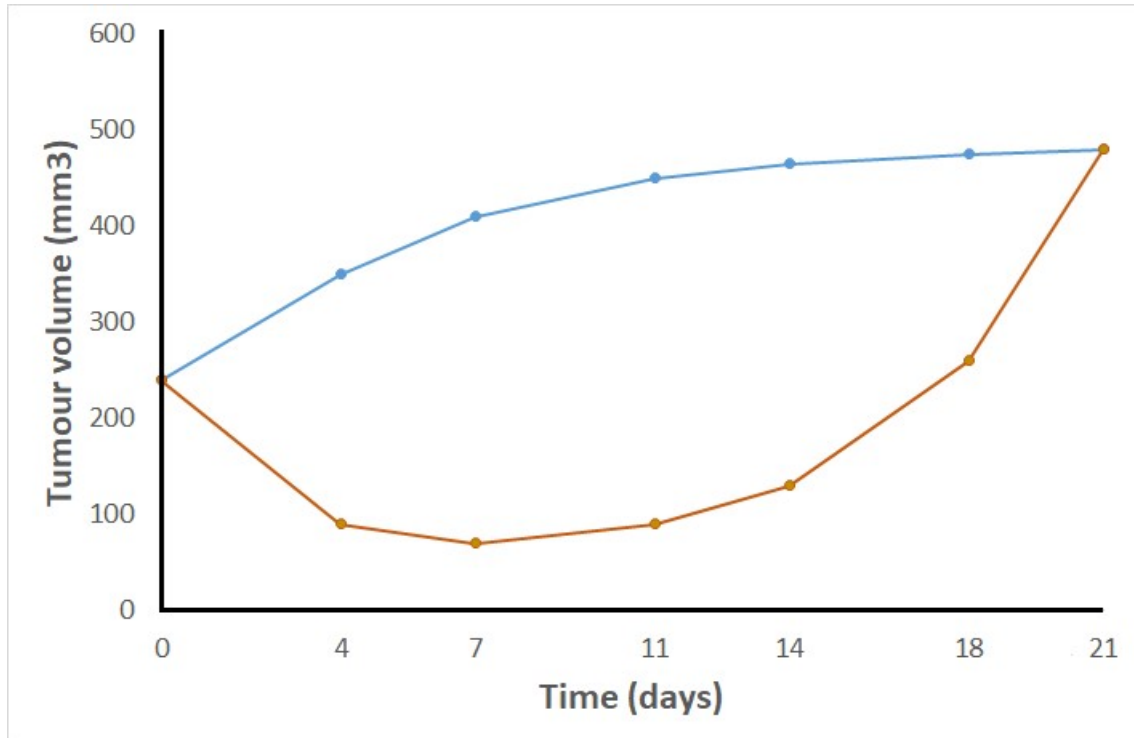


Figure 1.3.3.1 Repeated measures of tumor volume on two rodents, at the start of treatment until the doubling time. Tumor growth delay is the same but biological mechanisms are different

1.3.4 Multivariate methods

The methods are called “multivariate” because they treat the series of tumor volumes on an animal as a single multivariate observation (Heitjan *et al.*, 1993). They use the entire data series and permit detailed modelling of tumor growth curves and intra-animal correlation patterns, substantially improving the efficiency of testing and reducing sample size requirements. Furthermore, they provide more descriptive features that address mechanisms underlying tumor growth inhibition and maximize the biological information obtained from *in vivo* studies. Repeated-measures ANOVA, or Friedman repeated-measures ANOVA on ranks, can compare tumor growth curves after accounting for the correlation of measurements on the same tumor. Other multivariate models are reported in Heitjan *et al.* (1993). On the other hand, these multivariate methods could be criticized for the following reasons:

- in case normality or homoscedasticity (i.e. same variance in all groups and at all times) are assumed, they are often unreliable and, due to small samples, usually unverifiable
- in case a correlation structure between repeated measures is assumed, it is often unreliable and, due to small samples, usually unverifiable
- in case of missing values, either data series are excluded, or imputation is used, or a correlation structure should be specified. Due to informative missing and small samples, imputation techniques could introduce biases into the analyzed data.

Sophisticated regression models have been proposed to fit tumor growth curves; a biexponential model (Demidenko, 2004; Liang *et al.*, 2004), a linear exponential model (Demidenko, 2006), a non-parametric model (Liang, 2005) and a Bayesian model (Zhao *et al.*, 2011). However, regression models to fit tumor growth curves have limits: due to a small sample size of preclinical *in vivo* experiments, assumptions are only verifiable with great difficulty and, if an excessive number of parameters are used, overfitting occurs.

In addition to the statistical approaches reported in Sections 1.3.1-1.3.4, many statistical tests, unfortunately not combined with appropriate estimators (i.e. only p-values are obtained), have been proposed. Tan and colleagues (Tan *et al.*, 2002) proposed a small-sample t-test via the EM (expectation maximization) algorithm. They assumed a multivariate normal distribution for the repeated log tumor volumes with a Toeplitz covariance matrix. Due to the strong model assumption, their method has limited application to preclinical *in vivo* experiments. Vardi *et al.* (2001) proposed a nonparametric two-sample U-test. The proposed methodology is a fully nonparametric approach.

Finally, Liang (2007) proposed a non-parametric approach to compare antitumor effects in two treatment groups. The approach yields a p-value only.

In conclusion, different shortcomings are present in the current statistical methods used to analyse preclinical *in vivo* tumor growth curves: incomplete use of the entire data series, unreliable assumptions, poorly addressed biological mechanisms underlying different patterns of tumor growth, lack of statistical power and inferential estimators with inadequate statistical properties. This project would like to improve statistical methodology applied to preclinical *in vivo* tumor growth curves, overcoming previous shortcomings.

Chapter 2

Aims

There is no doubt that poorly designed, executed, analyzed and reported *in vivo* experiments raise ethical as well as scientific concerns. Briefly, on one side the weight of scientific evidence is reduced and no statistical method could completely fix this damage. On the other side, research reproducibility, the fundamental assumption of science, is definitely compromised. As a consequence of poor methodology applied to *in vivo* experiments, 'wrong roads' could be taken in preclinical research (Figure 2.1). Many laboratories spend time and money and use *in vivo* models in vain, trying to extend unreliable findings or apply them to different problems. The mean number of citations of the forty-seven landmark studies non-reproduced by the scientists in the Haematology and Oncology department at the biotechnology firm Amgen in Thousand Oaks, was about two hundred (range: 3-1.909 citations, Begley *et al.*, 2012) .



Figure 2.1 Consequence of poor methodology applied to *in vivo* experiments

It could be worse still. Methodological flaws in *in vivo* experiments could ruin the drug development process. It could be hypothesized that the recurring failure of apparently promising interventions to improve outcome in clinical trials has been partially caused by these flaws. For instance, several of these errors could have led to bias with false positive (i.e. type I) errors. And false positive errors could have wrongly selected CCs for clinical evaluation.

To the best of the author's knowledge, the impact of methodological flaws in *in vivo* experiments on the drug development process has never been quantitatively investigated. In Oncology, the assessment of antitumor activity in *in vivo* experiments and clinical trials could be a useful way to detect and estimate this impact. CCs demonstrating better antitumor activity than no treatment or standard therapies (i.e. active controls) in preclinical cancer models (i.e. *in silico*, *in*

vitro and *in vivo* models), are advanced to confirmatory testing in early (i.e. Phase I and II) clinical trials. Antitumor activity detected in preclinical cancer models is a fundamental prerequisite for advancing a CC from preclinical testing in the laboratory to clinical testing and for prioritizing CCs' progress to clinical cancer trials. This prerequisite is based on the assumption that CCs' activity in preclinical cancer models translates into at least some efficacy in human patients. In drug discovery and development, *in vivo* models have the greatest complexity and, above all, the greatest similarity to human patients among preclinical cancer models. At the same time, the assessment of antitumor activity is the first test bench of the CCs' clinical development after CCs' dose has been defined. Therefore, detecting and estimating the correlation between methodological quality of *in vivo* experiments, whose primary objective is to assess CCs' antitumor activity, and the level of CCs' antitumor activity in phase II clinical trials, could be a direct way to estimate the impact of methodological flaws in preclinical *in vivo* experiments on the process of drug discovery and development.

It is necessary to retrieve data about statistical design and analysis of preclinical *in vivo* tumor efficacy studies from scientific literature in order to address the previous issue. Hence, it is also possible to assess the methodological quality of *in vivo* tumor efficacy studies using the same data. To the best of the author's knowledge, the methodological quality of preclinical *in vivo* tumor efficacy studies has never been qualitatively and quantitatively investigated.

Finally, the statistical design and analysis of experiments to study *in vivo* tumor growth curves is a primary issue. A new methodological framework, based on the Wilcoxon-Mann-Whitney test, will be introduced for the statistical design and analysis of these experiments.

More specifically, the project addresses the following interrelated aims:

1. to correlate the quality of statistical design and analysis of preclinical *in vivo* tumor efficacy studies with the level of antitumor activity estimated in phase II clinical trials
2. to evaluate the quality of statistical design and analysis of preclinical *in vivo* tumor efficacy studies
3. to improve the statistical design and analysis of experiments to study tumor growth curves.

Regarding the first and second aim, research will focus on epithelial ovarian cancer (EOC). EOC was chosen as tumor type for the following reasons:

1. EOC treatment has not been substantially changed in the last thirty years. A platinum-based chemotherapy is the mandatory first line treatment. Stability and simplicity of administered treatment has a favourable influence on the control of variability

2. this project refers to the Oncology Department, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan (Italy). A large number of EOC research studies has been performed in this department in the last thirty years. Specifically, good expertise and skills has been developed in animal models and translational research about this type of tumor.

It was necessary to design a survey in order to achieve the first two aims of the project. This task has been difficult and time-consuming. Survey design has been profoundly amended. Two previous survey designs have been rejected because they could not be effectively applied. Their ineffective applicability was due to methodological limits of designs used for phase II trials in Oncology (e.g. phase II clinical trials in Oncology are generally single-arm trials) and publication bias (i.e. *in vivo* experiments are not clearly identifiable in public assessment reports published by the European Medicines Agency and the Food and Drug Administration). Failed survey designs will be described and discussed in Chapter 6.

Regarding the third aim, the project proposes a new methodological framework to study tumor growth curves. It is a general framework because it focuses on both statistical hypotheses testing and the theory of estimation.

Chapter 3

Methods

3.1 Survey design

To achieve the first two aims of the project, a systematic survey of previous clinical and preclinical research has been performed using a sequential two-stage design. In the first stage, eligible CCs have been identified and estimates of their antitumor activity has been retrieved from clinical research literature. In the second stage, *in vivo* experiments testing antitumor activity of identified CCs have been retrieved from biomedical research literature. The quality of experimental design and statistical analysis of each preclinical *in vivo* experiment has been evaluated using an ad hoc checklist. Finally, the quality of experimental design and statistical analysis of preclinical *in vivo* experiments has been correlated to the estimates of clinical antitumor activity. If methodological flaws impact the Oncology drug discovery and development process, a positive correlation between methodological quality of *in vivo* experiments and estimates of clinical antitumor activity could be expected.

Details about this two-stage design follows.

3.1.1 Stage 1: identification of eligible CCs and estimation of their clinical antitumor activity

A systematic search of the Medline and EMBASE databases has been carried out to identify clinical trials whose primary objective was to assess antitumor activity of CCs. This systematic search was limited to clinical trials in EOC. Reasons of this choice have been reported at the end of chapter 2.

Selection of EOC clinical trials and eligible CCs was based on the following criteria:

α. Eligible criteria for clinical trials

α1. Inclusion criteria

- ✓ Histologically or cytologically confirmed diagnosis of epithelial ovarian cancer, fallopian tube cancer, or primary peritoneal cancer
- ✓ Assessment of CCs' antitumor activity was the primary or co-primary objective. At least one antitumor activity endpoint was a primary or co-primary endpoint
- ✓ Women aged 18 years or older
- ✓ ECOG/WHO performance status (PS) 0-2 or GOG PS 0-2 (Oken *et al.*, 1982; Rubin *et al.*, 2004)
- ✓ Patients must have failed at least one prior line of platinum-based chemotherapy
- ✓ Study protocol approved by the independent ethics committees or institutional review boards of the participating institutions

- ✓ The final study report was published on 1st January 2010 or later
- ✓ The final study report was written in English

Note: if inclusion criteria of the EOC clinical trials were broader but all patients evaluated for antitumor activity satisfied all previous $\alpha 1$ criteria, the clinical trial was considered eligible. For instance, if eligible criteria admitted the enrollment of children and all effectively enrolled patients were adults, EOC clinical trial was considered eligible.

β. Eligible criteria for CCs

β1. Inclusion criteria

- ✓ CC was evaluated in monotherapy as experimental treatment (i.e. active arm)

β2. Exclusion criteria

- ✓ Monoclonal antibodies, oncolytic viruses or reoviruses, vaccines, immunotherapeutic and endocrine CCs were excluded
- ✓ CC was administered in maintenance therapy
- ✓ CC was administered as standard treatment (i.e. control arm)

The following search string was used in Medline:

("Clinical Trial, Phase II"[Publication Type] OR "clinical trial phase 2" OR "clinical trial phase ii" OR "clinical study phase 2" OR "clinical study phase ii" OR "phase 2 clinical study" OR "phase 2 clinical studies" OR "phase ii clinical study" OR "phase ii clinical studies" OR "phase 2 clinical trial" OR "phase 2 clinical trials" OR "phase ii clinical trial" OR "phase ii clinical trials" OR "phase 2 study" OR "phase 2 studies" OR "phase ii study" OR "phase ii studies" OR "phase 2 trial" OR "phase 2 trials" OR "phase ii trial" OR "phase ii trials" OR "phase 1/2" OR "phase 2/3" OR "phase 1 2" OR "phase 2 3" OR "phase 1-2" OR "phase 2-3" OR "phase i/ii" OR "phase ii/iii" OR "phase i ii" OR "phase ii iii" OR "phase i-ii" OR "phase ii-iii") AND ("Ovarian Neoplasms"[Mesh] OR "ovarian cancer" OR "ovarian cancers" OR "ovarian tumor" OR "ovarian tumors" OR "ovarian tumour" OR "ovarian tumours" OR "ovarian carcinoma" OR "ovarian carcinomas" OR "ovarian neoplasm" OR "ovarian neoplasms" OR "ovary cancer" OR "ovary cancers" OR "ovary tumor" OR "ovary tumors" OR "ovary tumour" OR "ovary tumours" OR "ovary carcinoma" OR "ovary carcinomas" OR "ovary neoplasm" OR "ovary neoplasms")

Filters: Publication date from 2010/01/01

The following search string was used in EMBASE:

#8 #6 AND #7

#7 [embase]/lim NOT [medline]/lim

#6 #3 AND #4 AND [2010-2019]/py

#5 #3 AND #4

'phase 2 clinical trial'/exp OR 'clinical trial phase 2' OR 'clinical trial phase ii' OR 'clinical study phase 2' OR 'clinical study phase ii' OR 'phase 2 clinical study' OR 'phase 2 clinical studies' OR 'phase ii clinical study' OR 'phase ii clinical studies' OR 'phase 2 clinical trial' OR 'phase 2 clinical trials' OR 'phase ii clinical trial' OR 'phase ii clinical trials' OR 'phase 2 study' OR 'phase 2 studies' OR 'phase ii study' OR 'phase ii studies' OR 'phase 2 trial' OR 'phase 2 trials' OR 'phase ii trial' OR 'phase ii trials' OR 'phase 1/2' OR 'phase 2/3' OR 'phase 1 2' OR 'phase 2 3' OR 'phase 1-2' OR 'phase 2-3' OR 'phase i/ii' OR 'phase ii/iii' OR 'phase i ii' OR 'phase ii iii' OR 'phase i-ii' OR 'phase ii-iii'

#3 #1 OR #2

'ovarian cancer' OR 'ovarian cancers' OR 'ovarian tumor' OR 'ovarian tumors' OR 'ovarian tumor' OR 'ovarian tumors' OR 'ovarian carcinoma' OR 'ovarian carcinomas' OR 'ovarian neoplasm' OR 'ovarian neoplasms' OR 'ovary cancer' OR 'ovary cancers' OR 'ovary tumor' OR 'ovary tumors' OR 'ovary tumor' OR 'ovary tumors' OR 'ovary carcinoma' OR 'ovary carcinomas' OR 'ovary neoplasm' OR 'ovary neoplasms'

#1 'ovary tumor'/exp

A PRISMA flow diagram (<http://prisma-statement.org/PRISMAStatement/FlowDiagram.aspx>. Retrieved 09 May 2019) was created to summarize the systematic review process.

After the identification of eligible EOC clinical trials, their following characteristics were retrieved:

- | | |
|---|---|
| ✓ CC evaluated | (e.g. sorafenib) |
| ✓ Phase of the clinical trial | (i.e. phase I/II; II; II/III) |
| ✓ Year of final report publication | (e.g. 2012) |
| ✓ Participating macro geographical regions | (i.e. USA; Europe; other countries) |
| ✓ Monocentric or multicenter trial | |
| ✓ Single arm or controlled trial | |
| ✓ Randomized trial, in case of controlled trial | (i.e. yes; no) |
| ✓ Open-label or blind trial | (e.g. double-blind trial) |
| ✓ Start date and end date of recruitment | (e.g. October 2013) |
| ✓ Major eligibility criteria | (e.g. number of previous treatment lines) |
| ✓ Demographic and pathological characteristics of enrolled patients (e.g. age; histotype) | |
| ✓ Characteristics of administered treatments | (i.e. schedule) |

- ✓ Primary activity endpoint (e.g. Disease Control Rate)
- ✓ Tumor response criteria (e.g. RECIST, version 1.0)
- ✓ Blinded Independent Central Review of tumor response (i.e yes; no)
- ✓ Objective Response Rate (ORR): point estimates and 95% CI, overall and by platinum-sensitivity
- ✓ Median progression-free survival (PFS): point estimates and 95% CI, overall and by platinum-sensitivity

3.1.2 Stage 2: systematic review of *in vivo* experiments and their methodological evaluation

For each eligible CC, whose clinical antitumor activity has been estimated in stage 1, a systematic search of Medline and EMBASE databases was carried out to identify scientific papers reporting original antitumor activity research on live rats and mice. These rodents are the most widely used animals in antitumor activity studies. If other species were used, their inclusion would reduce the sensitivity of the survey for drawing inferences about quality of statistical design and analysis.

Selection of *in vivo* experiments about each eligible CC was based on the following criteria:

γ. Eligible biomedical research papers

γ1. Inclusion criteria

- ✓ At least one *in vivo* experiment on mouse or rat models of EOC model was reported
- ✓ Eligible CC was used in *in vivo* experiment on mouse or rat models of EOC
- ✓ The study report was written in English

δ. Eligible *in vivo* experiment on mouse or rat models of cancer

δ1. Inclusion criteria

- ✓ The primary or co-primary objective was to assess antitumor activity of CC
- ✓ eligible CC was evaluated in monotherapy

δ2. Exclusion criteria

- ✓ *In vivo* pharmacokinetics or toxicology experiments
- ✓ Pharmacodynamics experiments

A PRISMA flow diagram (<http://prisma-statement.org/PRISMAStatement/FlowDiagram.aspx>. Retrieved 09 May 2019) was created to summarize the systematic review process.

Search strings used in Medline and EMBASE are reported in Appendix A. These strings were developed exploring four primary concepts: epithelial ovarian tumor, preclinical screening, rat or mouse species and CC's name.

After the identification of eligible biomedical research papers, their following characteristics were retrieved:

- ✓ Year of publication (e.g. 2000)
- ✓ CC evaluated (e.g. Sorafenib)
- ✓ Peer-reviewed journal (i.e. yes; no)
- ✓ Conflict of interest (i.e. declared; *not declared*)
- ✓ Funding statement (i.e. private; public; mixed; *not declared*)
- ✓ Recommended clinical translation (i.e. yes; no)

After the identification of eligible *in vivo* tumor efficacy experiments, the following characteristics of each mouse or rat cancer model were retrieved:

- ✓ Species (i.e. mouse; rat; *not declared*)
- ✓ Strain (e.g. BALB/c Nude; *not declared*)
- ✓ Genotype (i.e. inbred; outbred; *not declared*)
- ✓ Immune status (i.e. competent; compromised; *not declared*)
- ✓ Sex (i.e. female; male; mixed; *not declared*)
- ✓ Age (weeks) (e.g. 5-7 wks; *not declared*)
- ✓ Weight (grams) (e.g. 17 grams on average; *not declared*)
- ✓ Housing (i.e. SPF; Conventional; *not declared*)
- ✓ Animal source (i.e. internal breeding; certified breeder; *not declared*)
- ✓ Model type (i.e. human xenograft PDX; human xenograft CDX; syngeneic; genetically engineered; other, specify; *not declared*)
- ✓ Human ovarian cancer cell line (e.g. SKOV3)

Next, for each eligible *in vivo* tumor efficacy experiment, the quality of statistical design and analysis was assessed using the checklist described in the following section.

3.2 Quality of statistical design and analysis checklist

The checklist used to assess the quality of statistical design and analysis of *in vivo* tumor efficacy experiments is reported in Appendix B. The checklist consists of 46 items separated in 8 distinct sections. Separate sections are related to distinct methodological issues. The references Heitjan *et al.*, 1993; Altman *et al.*, 1995; Kilkenny *et al.*, 2009; Kilkenny *et al.*, 2010; Hooijmans *et al.*, 2014; Henderson *et al.*, 2015, were used to draft and finalise the checklist. The checklist has been tested in a small pilot study of 20 eligible papers. The judgement has been subjective. If the checklist missed something important in one of the twenty papers, the checklist was updated.

The contents of the checklist's sections are summarised below.

3.2.1 Repetition and external validity

On one side, the repetition of the same animal experiment by the same laboratory setting is ethically and scientifically unsound. If a laboratory wants to repeat its own experiment, it probably means it was badly designed (e.g. underpowered study) or other methodological errors (e.g. lack of well-defined and standard operating procedures) were performed. On the other side, evidence that leads to a single laboratory or a single model or species showing some benefit should not be used as the basis for proceeding to the clinic (van der Worp et al., 2010). Hence, for each eligible *in vivo* experiment, the following information was retrieved: the number of repetitions by a single laboratory, the number of species and models used for the EOC malignancy type, and the number of participating laboratories.

3.2.2 Internal validity

As shown in Chapter 1, internal validity is at the core of good experimentation. To prevent selection bias, the use of an internal control group, the random allocation of animals, and methods used to conceal the allocation sequence were recorded. Control of variability is fundamental in *in vivo* experiments. Russell and Burch's chapter on reduction, that was written back in 1959, stressed the use of inbred strains to control inter-individual variation (Russell *et al.*, 1992). Poor control of variability increases the probability of false positive and false negative results (Kernan *et al.*, 1999). Hence, for each eligible *in vivo* experiment, type of randomization (i.e. simple, block, stratified and unequal randomization) and the use of randomization techniques for other sampling units other than individual animals, such as cages, was recorded. Correct identification of the experimental unit [i.e. the unit that is randomly assigned to a treatment (Casella, 2008)] and cases, where there was suspicion of pseudo-replication, was recorded. Finally, it was checked whether the CC monotherapy arm was primarily used as active or control comparator arm, for each eligible *in vivo* experiment. However, this was only for descriptive purposes. Whereas the survey of clinical trials was limited to those in which the CC was the active-treatment arm, in the survey of *in vivo* experiments the CC could be either arm.

3.2.3 Statistical design

As Casella G. taught brilliantly to us, there are two aspects to a design: treatment design and experiment design (Casella, 2008). A statistical design contains both. A treatment design is the manner in which the levels of treatment factors are arranged in an experiment. Typically, treatment

factors are either crossed or nested (refer to Figure 3.2.3.1 and 3.2.3.2), and this relationship can be either complete or incomplete.

First treatment factor	A											
Level	1			2			3			4		
Second treatment factor	B			B			B			B		
Level	1	2	3	1	2	3	1	2	3	1	2	3
Treatment group	1	2	3	4	5	6	7	8	9	10	11	12

Figure 3.2.3.1 Treatment factor A and treatment factor B are completely crossed

First treatment factor	A											
Level	1			2			3			4		
Second treatment factor	B			B			B			B		
Level	1	2	3	4	5	6	7	8	9	10	11	12
Treatment group	1	2	3	4	5	6	7	8	9	10	11	12

Figure 3.2.3.2 Treatment factor A and treatment factor B are completely nested

For the first aspect of a statistical design (i.e. treatment design), the use of a factorial (i.e. treatment factors are completely or partially crossed) and dose-response (i.e. ≥ 3 doses) design was recorded. It was checked if an active control group was used. The number of design factors (i.e. first and second treatment factors in Figure 3.2.3.1 and 3.2.3.2) and treatment groups were also recorded.

An experiment design determines the way in which the randomization of experimental units to treatments is carried out and how the data are actually collected. The error structure of the experiment is a consequence of the experiment design. Examples of different experiment designs are: the completely randomized design, in which all the treatment combinations of factors A and B are randomly allocated to units throughout the design; the randomized complete block design, in which all the treatment combinations of factors A and B are randomly allocated to units within blocks; the strip plot design, in which the randomization of both treatment factors are restricted within block factors. In Figure 3.2.3.3 possible “field layout” for completely randomized (a), randomized complete block (b), and strip plot designs (c) are shown.

(a) Whole experiment with 2 reps shown					
A1B1	A1B1	A2B3	A2B1	A3B2	A1B3
A1B2	A1B2	A3B2	A2B1	A2B2	A3B3
A3B1	A2B2	A2B3	A3B1	A3B3	A1B3

(b) One block shown			(c) One complete block shown		
A2B1	A3B2	A1B3	A1B1	A1B2	A1B3
A1B1	A2B2	A3B3	A3B1	A3B2	A3B3
A3B1	A1B2	A2B3	A2B1	A2B2	A2B3

Figure 3.2.3.3 Possible layouts for three different experiment designs

In the completely randomized design (a), neither of the two rectangles of 9 units contains all combinations of A and B, so they are not complete blocks. In the randomized complete block design (b), 9 units per block are necessary to accommodate all combinations of A and B. In the strip plot design (c), the two block factors are rows and columns. All the designs, (a), (b) and (c), will then allow to estimate and test the main effects of A, the main effects of B, and the $A \times B$ interaction effects.

For the second aspect of a statistical design (i.e. experiment design), it was checked if blocking was used and whether the experimental unit was assigned to more than one treatment group (i.e. crossover design or within units design).

3.2.4 Sample size

In *in vivo* experiments sample size is a critical factor in detecting a relevant biological signal, avoiding false negative results and observing ethical requirements. As the ANOVA framework teaches us, replication (i.e. the repetition of the experimental situation by replicating the experimental unit) is necessary in order that the biological signal prevails on the error structure of the experiment. From a scientific point of view, underpowered studies should be avoided, as they might lead to the false conclusion that the CC is ineffective and all included animals will have been used to no benefit. From an ethical point of view, overpowered studies should be avoided, as animals are wasted. For each *in vivo* experiment, it was checked if the sample size was justified and which method was used to justify the sample size. The number of animals per treatment group declared in the methods section of the eligible paper was also recorded.

3.2.5 Outcomes and their assessment

First of all, it was checked whether the primary outcomes were clearly identified in the methods section of the eligible paper. Then the list of primary and secondary outcomes and how they were defined was recorded. As task difficulty was considered very high and results unreliable due to poor reporting of preclinical *in vivo* experiments, selective outcome reporting was not evaluated in this survey.

In *in vivo* antitumor efficacy studies, tumor growth is followed from tumor inoculation to a particular event of interest, that is called the antitumor activity endpoint (e.g. tumors reach a predetermined target volume). However, the event of interest may not be observed for some animals because of end-of-study censoring or competing events, such as death due to toxic effects of antitumor therapies or animal sacrifice that is ethically necessary. If the antitumor activity endpoint or the competing events are not clearly defined, the interpretation of the tumor growth curve could be misleading. For example, a relatively slower growth rate could be due either to effect of treatment or to exclusion of animals that have been sacrificed because of a competing event. It was checked whether the antitumor activity endpoint and all the competing events were clearly defined.

Then, it was recorded whether the caregivers, investigators and outcome assessors were blinded from knowledge which intervention each animal received during the experiment. Finally, to prevent detection bias, it was recorded whether animals were selected at random for outcome assessment.

3.2.6 Statistical analysis

Data analysis of *in vivo* antitumor efficacy studies was assessed in this section. First of all, it was checked whether inferential methods were used to demonstrate antitumor activity. Inferential methods were classified as hypothesis testing or estimation methods. If used, the inferential methods were recorded. If applicable, it was checked whether statistical assumptions used to analyze tumor growth data were justified and whether methods for correction of multiple comparison were used. Finally, descriptive methods used to demonstrate antitumor activity were recorded.

3.2.7 Attrition bias about tumor growth curves

Incomplete outcome data causes ambiguity in the interpretation of results. The number of animals assigned to each treatment arm and reported in the results section was recorded. It was checked whether and how many animals, assigned to each treatment arm, were excluded from statistical

analysis. It was checked whether the number of animals with right-censored data was clearly reported and, more specifically, whether, for each animal included in the statistical analysis, it was clearly reported which event determined the end of follow-up. In clinical trials, it is a common practice to report progressively the number of patients at risk in the plot of survival curves. It was checked, if applicable, whether the number of animals at risk were progressively reported in the plot of tumor growth or survival curves. It is not enough to report progressively descriptive or inferential measures of variability such as standard deviation or standard error, respectively. A correct interpretation of these measures needs to specify the number of animals to which these measures refer. Finally, it was checked whether the length of follow-up was clearly defined. Treatments effect varies greatly along time; for example, a cytotoxic CC could have an effect that is very different from a cytostatic CC, not only in its magnitude, but also in its pattern over time. Hence, the comparison of tumor growth or survival curves depends heavily on the length of follow-up chosen.

3.2.8 Miscellanea

The majority of statistical analyses are performed with the help of computer programs. It was checked if any information was given in the papers about the commercial software used to analyse the data. Further, to evaluate whether known statistical involvement improved the quality of a paper, it was checked if any author was a member of a department of statistics or epidemiology. Finally, any peculiarities which the assessor noticed in the papers and which were not covered in other sections of the checklist were recorded.

3.3 Sample size

The sample size was not based on formal statistical considerations. It was determined by the time period in which final reports of eligible phase II clinical trials were published. This time period had to satisfy the following criteria:

1. it was the most recent time period to the project execution
2. large enough to retrieve 50 to 100 eligible phase II clinical trials from clinical research literature.

This upper limit was considered satisfactory in order to assess and extract information from each biomedical publication within the planned project timeframe.

3.4 Statistical analysis

Non-parametric statistics (i.e. median and range for continuous variables, absolute and percentage frequencies for categorical variables) were used to describe eligible clinical trials and *in vivo* experiments. Antitumor activity of the eligible CCs was assessed in terms of ORR and median PFS. Separate analyses were performed for ORR and PFS endpoints. For each eligible CC identified in a specific cohort of patients, a point estimate and a standard error of ORR and median PFS was obtained. Point estimates and standard errors were then used as inputs for a random-effects meta-analysis. The Q and I² statistics were used to test and estimate, respectively, the percentage of total variation due to inter-cohort heterogeneity.

For the ORR endpoint, in order to include cohorts with an estimated proportion equal to 0 or 1, and since the coverage probability is closer to the nominal confidence level than that obtained from the exact likelihood approach, score test-based confidence intervals were used to estimate the percentage of patients with complete or partial response. The Stata *metaprop* command with *random*, *cimethod(score)* and *ftt* options was used (Nyaga *et al.*, 2014).

For the PFS endpoint, assuming an exponential distribution, the constant hazard rate was directly computed from the point estimate of the median PFS. The following formula was used:

$$\log(2) / \text{median PFS} \quad 3.4.1$$

Standard error of the constant hazard rate was estimated using the following formula:

$$\{ [\log(2) / \text{lower 95\%CI median PFS}] - [\log(2) / \text{upper 95\%CI median PFS}] \} / [2 \cdot \Phi(0.975)] \quad 3.4.2$$

where $\Phi(z)$ is the Standard Normal Cumulative Distribution Function. The Stata *metan* command with *random* option was used to estimate the random effects model (Harris *et al.*, 2010).

The checklist used to assess the quality of statistical design and analysis of *in vivo* tumor efficacy experiments was analysed by item, using non-parametric statistics (i.e. median and range for continuous variables, absolute and percentage frequencies for categorical variables). The unit of analysis was each eligible *in vivo* experiment.

A meta-regression approach was used to detect correlation between methodological quality of *in vivo* experiments and clinical antitumor activity.

The following procedure was applied:

- in the presence of multiple cohorts evaluating the same CC, a summary measure of the CC's clinical antitumor activity was estimated using a fixed-effects meta-analysis model
- *in vivo* experiments were classified as either experiments of good methodological quality ('good' *in vivo* experiment) or bad methodological quality ('bad' *in vivo* experiments). *In vivo*

experiments were considered of good methodological quality if the following checklist's items were positively answered: $\beta 1$, $\beta 1 a$, $\delta 1$, $\epsilon 1$, $\epsilon 3$, $\epsilon 5 c$, $\zeta 1$, $\eta 5$ and $\eta 6$ (refer to Appendix B for the items specification). If at least one of these items was negatively answered, the *in vivo* experiment was classified as a '*bad*' *in vivo* experiment

- based on the number of their '*good*' and '*bad*' *in vivo* experiments, CCs were ordered and weighted
- correlation between ordered CCs and their clinical antitumor activity was detected using a linear meta-regression model and was shown using a forest plot.

Statistical analyses were generated using SAS software, version 9.4 of the SAS System for Windows. Copyright (c) 2016 by SAS Institute Inc., Cary, NC, USA. Fixed and random effects model were generated using StataCorp. 2017. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC. Forest plots were generated using SAS software, version 9.4, and Review Manager (RevMan), version 5.3. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration, 2014.

Chapter 4

Results

4.1 Selection of clinical trials and CCs

On 1st January 2019 a systematic search of the Medline and EMBASE databases was carried out to identify eligible CCs and to estimate their clinical antitumor activity. The systematic database search yielded 2020 records. After 70 duplicates were removed, 1738 were excluded after reviewing the title and abstract. A total of 212 articles were selected for full-text review and closer inspection to determine whether they met eligibility criteria for clinical trials and CCs. 143 full-text articles were excluded, the major reasons being (1) duplicate abstract of an eligible and assessed trial ($n = 44$, 31%), (2) on-going eligible trial ($n = 31$, 22%), (3) ineligible CC or therapy ($n = 30$, 21%) or antitumor activity endpoint was neither primary nor co-primary of the trial ($n = 17$, 12%; refer to Figure 4.1.1). Sixty-nine eligible clinical trials in total were included in this survey. They are listed in Appendix C.

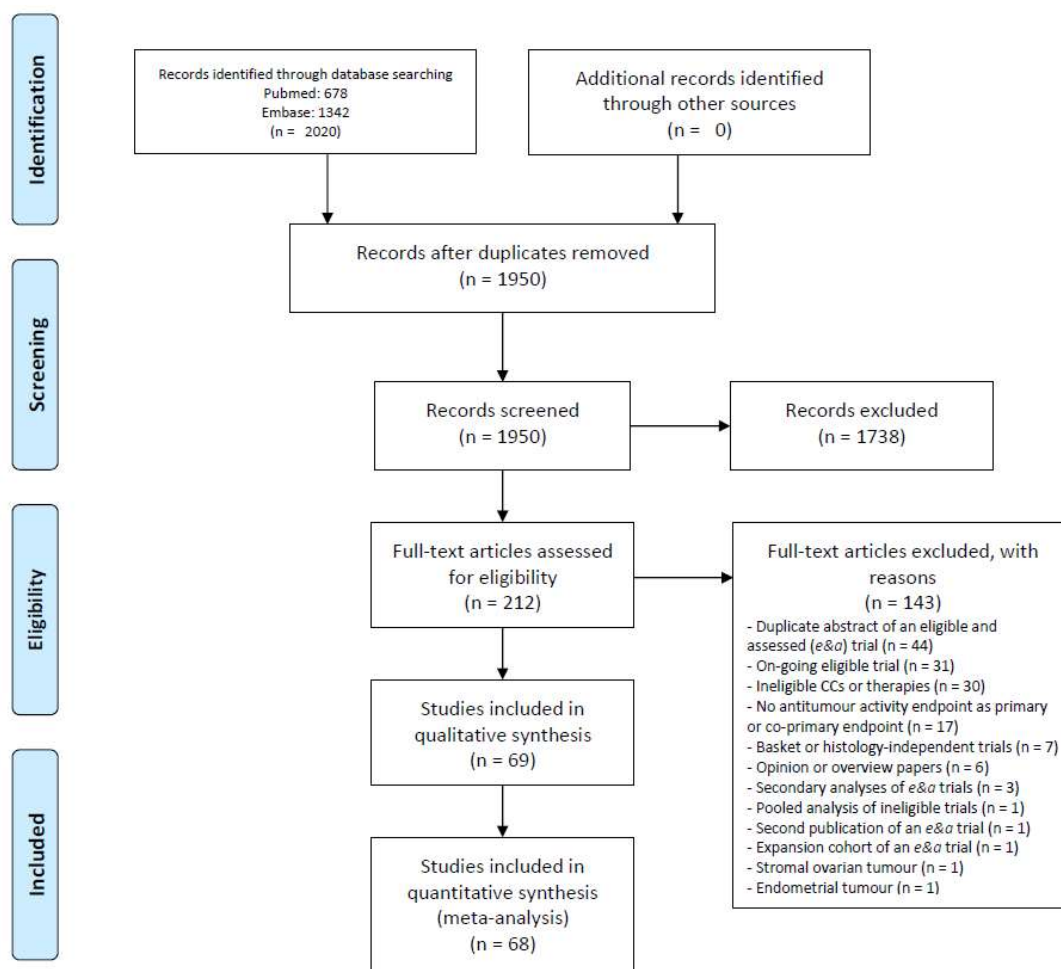


Figure 4.1.1 PRISMA flow diagram on selection of clinical trials

The major characteristics of eligible clinical trials are reported in Table 4.1.1. All the clinical trials except one (Tew *et al.*, 2014; see Appendix C) did not use blinding. Fifty-seven out of 69 (83%) were open-label, single-arm phase I/II or II trials. Only 5 out of 69 (7%) trials were randomized in conjunction with a control arm to detect efficacy improvement of the experimental CC with respect to the standard therapy. All the clinical trials except one (Seetharamu *et al.*, 2010; see Appendix C) started patients' accrual in the 21st century. Eligible CCs were evaluated as monotherapy in 85 cohorts of patients. In each cohort, the median number of patients evaluated for antitumor activity was 34.5 (IQR: 20.5-51.0). 49 out of 69 (71%) trials chose ORR as the primary endpoint. To assess tumor response, RECIST criteria, version 1.0 and version 1.1, were used respectively in 37 (54%) and 27 (40%) out of 68 trials (Graziani *et al.*, 2017, did not report which tumor response criteria were used; see Appendix C). GCIG CA125 criteria alone or in combination with RECIST criteria were used in 14 out of 68 (21%) trials. Sixty-four out of 69 (93%) trials did not use an Independent Review Committee to assess tumor response.

		<i>N</i>	%
Year of final publication			
2010-2012		30	43
2013-2015		16	23
2016-2019 ^{\$}		23	33
Phase of clinical trial			
II		65	94
I/II		4	6
Experimental design			
Blinding	Open-label	68	99
	Double-blind	1	1
Single-arm	Yes	57	83
	No	12	17
Randomized trial	Yes	12	17
	Controlled	5	42
	Not controlled, multi-arm	7	58
	No	57	83

Multicenter	Yes	58	84
	No	11	16
Geographic location			
USA		42	61
Europe		28	41
Other location	Canada	11	16
	Australia	5	7
	Israel	3	4
	Japan	2	3
	Brazil, China, South Korea	1	1
Start of accrual, year			
Median		2008	
Min-max		1999-2015	
<i>Missing data</i>		2	3
Efficacy population, number of patients by cohort ^{§§}			
N		84	
Median		34.5	
Q1-Q3		20.5-51.0	
Min-max		4-193	
<i>Missing data ^{§§§}</i>		1	1
Primary endpoint			
ORR		40	58
PFS		15	22
PFS AND ORR		9	13
DCR		5	7

Tumor response criteria		
RECIST Version 1.0	31	45
RECIST Version 1.1	20	29
RECIST Version 1.1 AND/OR GCIG CA125 criteria	7	10
RECIST Version 1.0 AND/OR GCIG CA125 criteria	6	9
WHO criteria	2	3
MASS criteria (Smith <i>et al.</i> , 2010)	1	1
GCIG CA125 criteria	1	1
<i>Missing data</i>	<i>1</i>	<i>1</i>
Tumor response assessment		
Investigator-determined response	64	93
Unblinded Independent Review Committee	5	7

Table 4.1.1 Major characteristics of eligible clinical trials

Legend 4.1.1: § Two articles in press on 1st January 2019 were considered eligible; \$\$ antitumor activity was evaluated in 85 cohorts of patients analyzed in the 69 eligible clinical trials; \$\$\$ for one cohort, the number of patients evaluated for antitumor activity was not available (Drew *et al.*, 2016; see Appendix C)

Fifty-two eligible CCs were identified in the 69 eligible clinical trials. They are listed in Table 4.1.2. Of these, 34 (65%) and 18 (35%) could be broadly classified as targeted and chemotherapeutic CCs, respectively. Forty-three out of 52 (83%) CCs were identified in a single clinical trial.

Classification of eligible CCs	CCs' names (number of eligible clinical trials in which the CC was administered as monotherapy)
Targeted therapy	Olaparib (4), Sunitinib (4), Cabozantinib (3), Rucaparib camsylate (3), Sorafenib (3), ENMD-2076 (2), Temsirolimus (2), Veliparib (2), Aflibercept (1), Alisertib (1), Apatinib (1), Birinapant (1), BI 2536 (1), Cediranib (1), Dalantercept (1), Danusertib (1), Dasatinib (1), Enzastaurin (1), Iniparib (1), Imatinib Mesylate (1), Lapatinib (1), L-asparaginase (1), Lenalidomide (1), Motesanib (1), Nintedanib (1), Pazopanib (1), Perifosine (1), Prexasertib (1), RO4929097 (1), Selumetinib (1), Tasquinimod (1), Urokinase-derived peptide (A6) (1), Vandetanib (1), Volasertib (1)

Chemotherapy	Topotecan (3), Belinostat (1), Bendamustine Hydrochloride (1), Elacytarabine (1), Eribulin Mesylate (1), Etririnotecan pegol (1), Gimathecane (1), Irofulven (1), Ixabepilone (1), Liposomal cisplatin (1), Lurbinectedin (1), Nab-paclitaxel (Abraxane) (1), Non-pegylated liposomal doxorubicin (1), Paclitaxel in non-protein lipid core nanoparticles (1), Patupilone (1), Sagopilone (1), Trabectedin (1), Zoptarelin Doxorubicin Acetate (1)
--------------	--

Table 4.1.2 List and broad classification of eligible CCs

4.2 Patient characteristics and assessment of clinical antitumor activity

One out of the 69 (1.4%) eligible trials was excluded from quantitative analysis of clinical antitumor activity because neither ORR nor PFS endpoints were available (Drew *et al.*, 2016). A total of 3455 patients were enrolled in the remaining 68 trials. The major characteristics of these patients are reported in Table 4.2.1. The mean age of patients was 60 years. The majority of enrolled patients had ECOG PS equal to 0-1. Despite the high percentage of missing data, it could be said that ovary and serous were the most common primary sites of tumor origin and cell type, respectively. Patients were mainly platinum resistant. Platinum resistant patients are those in whom the disease has progressed during (i.e. more precisely, refractory patients) or within 6 months of completing a platinum-based therapy. Platinum-sensitive are those with a platinum-free interval of six months or longer. At least 1547 (67%) patients had more than the one line of chemotherapy for recurrent disease.

	N	%
Age (years)		
N	3455	
Mean (Standard error)	60 (0.4)	
ECOG Performance Status		
0-1	3002	96
2	112	4
Missing data	341	10

Malignancy		
Ovarian cancer	1933	89
Primary peritoneal carcinoma	183	8
Fallopian tube carcinoma	58	3
<i>Missing data</i>	<i>1281</i>	<i>37</i>
Cell type		
Serous	2060	79
Adenocarcinoma	132	5
Endometrioid	126	5
Clear cell	120	5
Other	186	7
<i>Missing data</i>	<i>831</i>	<i>24</i>
No. of prior chemotherapy regimens		
1	760	33
>1	1547	67
<i>Missing data</i>	<i>1148</i>	<i>33</i>
Platinum sensitivity [§]		
Resistant	2064	65
Sensitive	1095	35
<i>Missing data</i>	<i>296</i>	<i>9</i>

Table 4.2.1 Major characteristics of enrolled patients

Legend 4.2.1: [§] Platinum resistant patients are those in whom disease has progressed during (i.e. refractory) or within 6 months of completing a platinum-based therapy

Antitumor activity of the 52 eligible CCs was assessed in terms of ORR and median PFS endpoints, in 82 and 72 cohorts, respectively. Statistical analysis of these endpoints is graphically shown in Figure 4.2.1 and Figure 4.2.2 respectively, and could be summarized by the following considerations:

- Most patients had a poor prognosis. Average summary ORR and median PFS were 13% (95%CI 10-16) and 3.3 (95%CI 2.9-3.9) months, respectively. These average estimates could not be attributable to the presence of few cohorts with very unsatisfactory prognosis. In fact, as shown in Figure 4.2.1 and Figure 4.2.2, these average estimates were obtained with an

approximately equal weight assigned to each cohort. ORR was less than 10% in 42 out of 82 (51%) cohorts and median PFS was less than 4 months in 40 out of 72 (56%) cohorts

- Cohorts' prognosis was highly heterogeneous. The proportion of variance not explained by random error was 85.9% and 95.7% for the ORR and median PFS endpoints, respectively. Although it is impossible to estimate the proportion of this unexplained variance due to heterogeneous distribution of prognostic and predictive factors, the great contribution of prognostic and predictive factors was shown by a couple of subgroup-analyses. In the first subgroup-analysis, the impact of platinum sensitivity on patients prognosis was estimated (refer to Figure 4.2.3 and 4.2.4). Cohorts evaluating platinum sensitive patients had a better prognosis compared to cohorts with platinum resistant patients, in terms of both ORR [average ORR: 31% (95%CI 23-41%) vs 10%(95%CI 7-14%), p-value<0.001] and median PFS [average median PFS: 5.8 (95%CI 4.8-7.4) vs 3.2 (95%CI 2.8-3.8) months, p-value<0.001]. However, platinum sensitivity did not fully account for the variance not explained by random error, suggesting that many, some of them unknown, prognostic/predictive factors cause the detected heterogeneity. In the second subgroup-analysis, CCs administered in more than one cohort were identified. In each subgroup of cohorts, the proportion of variance not explained by random error was estimated (refer to Table 4.2.2). Although the same CC was administered in the cohorts' subgroups, the cohorts' prognosis continued to be highly heterogeneous. However, there were exceptions to this, as clearly shown in Table 4.2.2.

In Figure 4.2.5 estimates of treatment effects in all eligible randomized and controlled clinical trials (RCTs) are shown. Because prognostic factors do not contribute to these estimates, the high heterogeneity between treatment effects observed in this group of trials supports our objective of correlating preclinical methodological quality to estimates of clinical antitumor activity. But the sample of available RCTs is too small to perform any meaningful analysis.

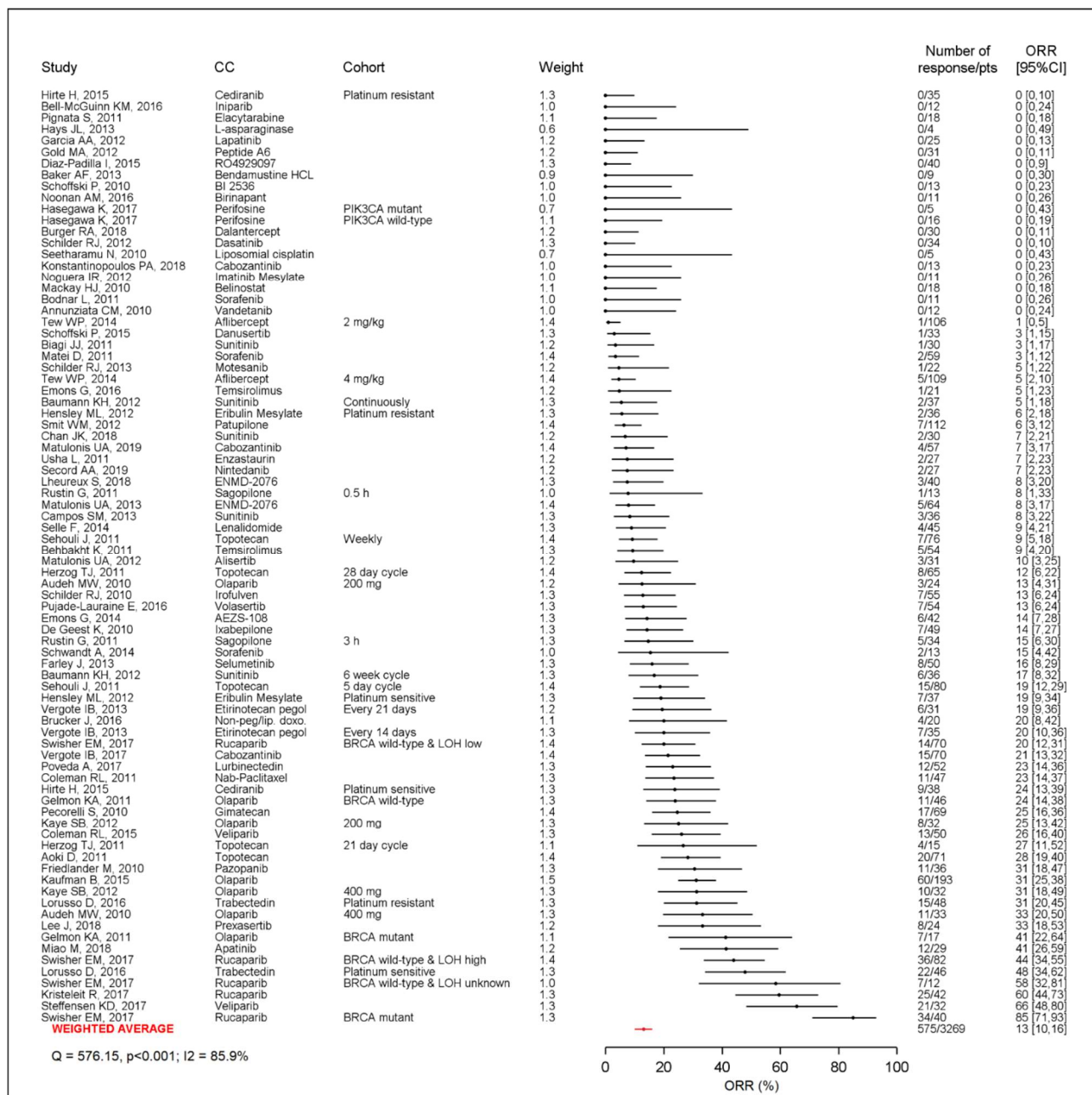


Figure 4.2.1 ORR distribution

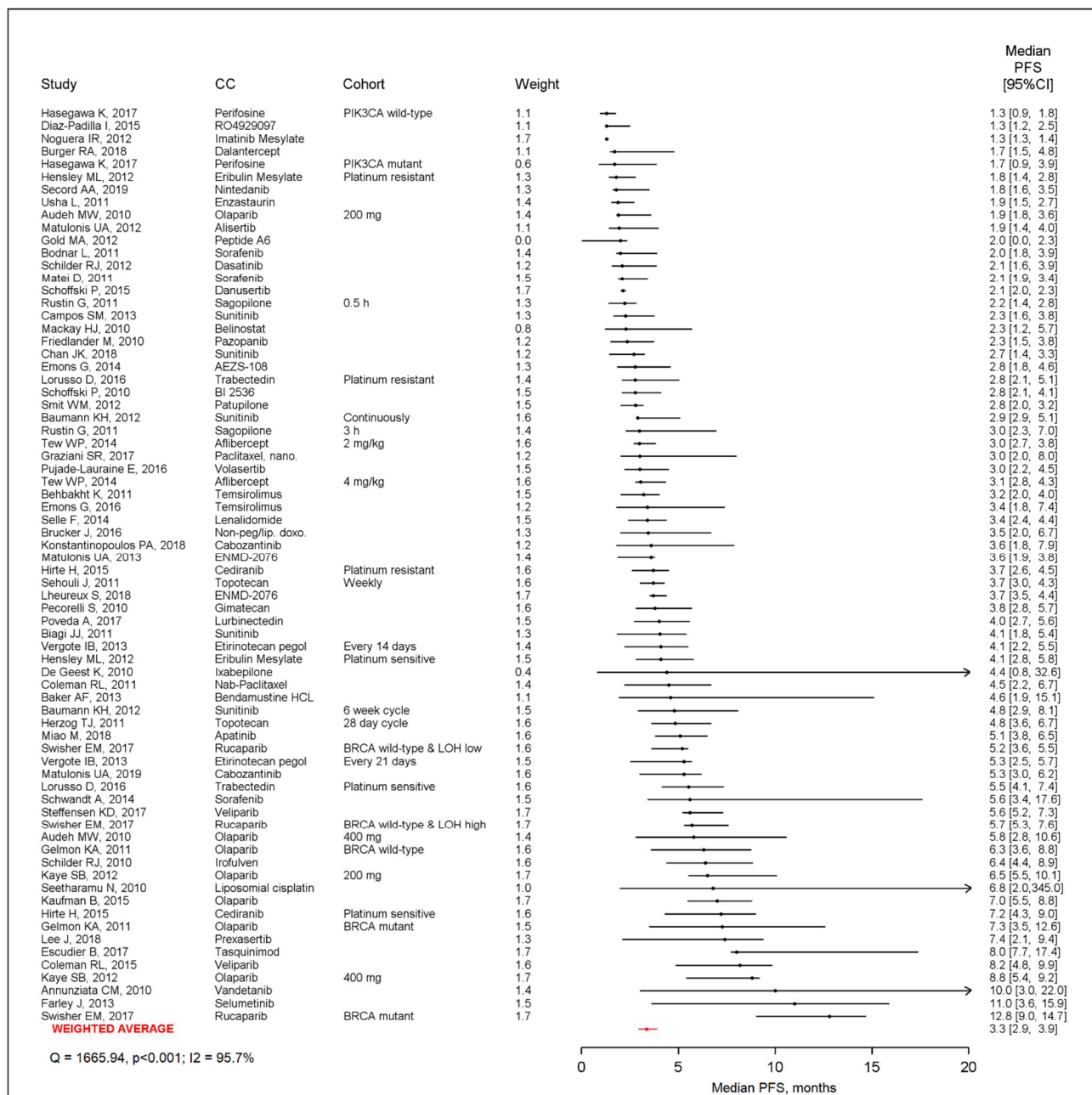


Figure 4.2.2 Median PFS distribution

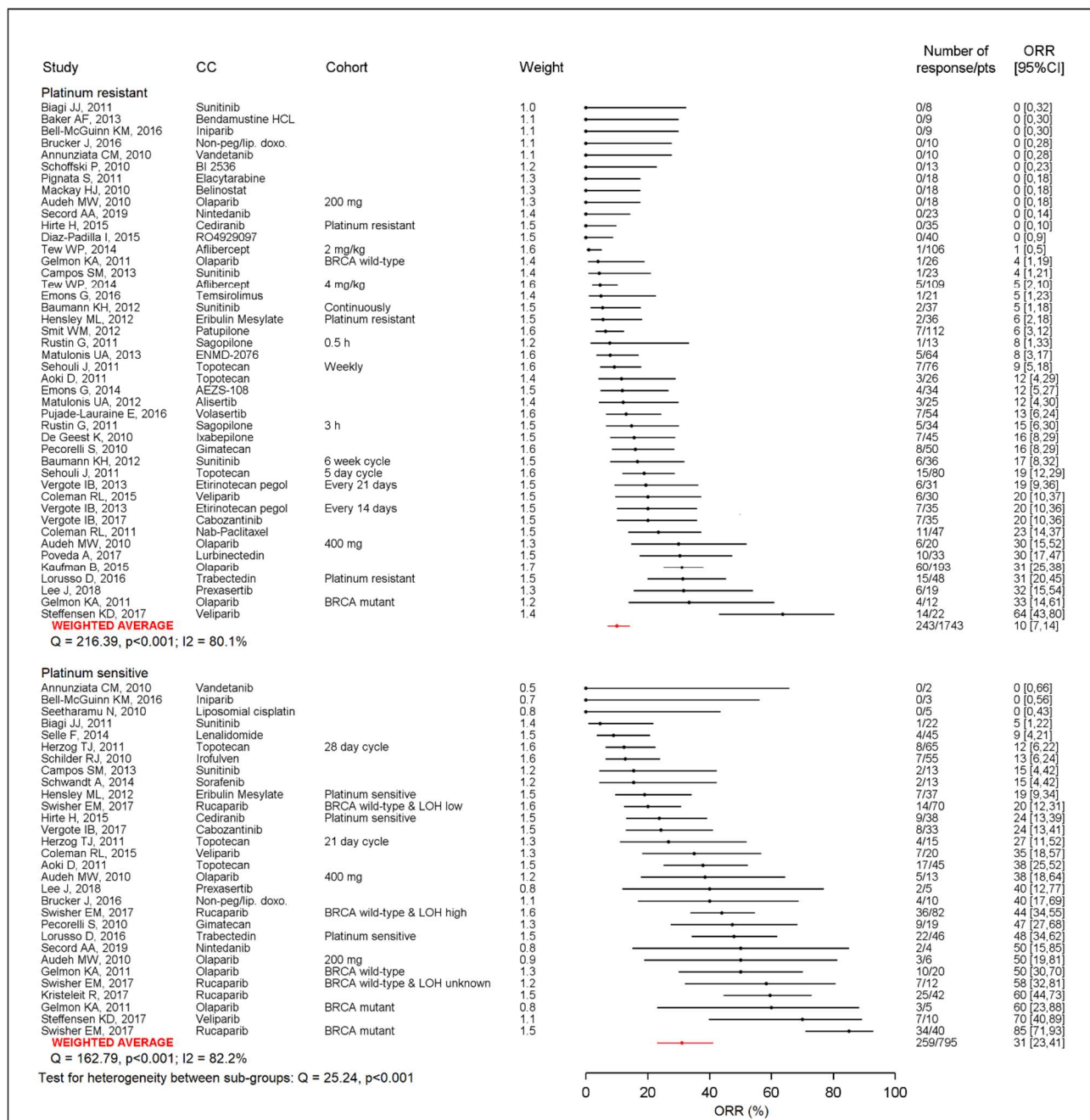


Figure 4.2.3 ORR distribution by platinum sensitivity

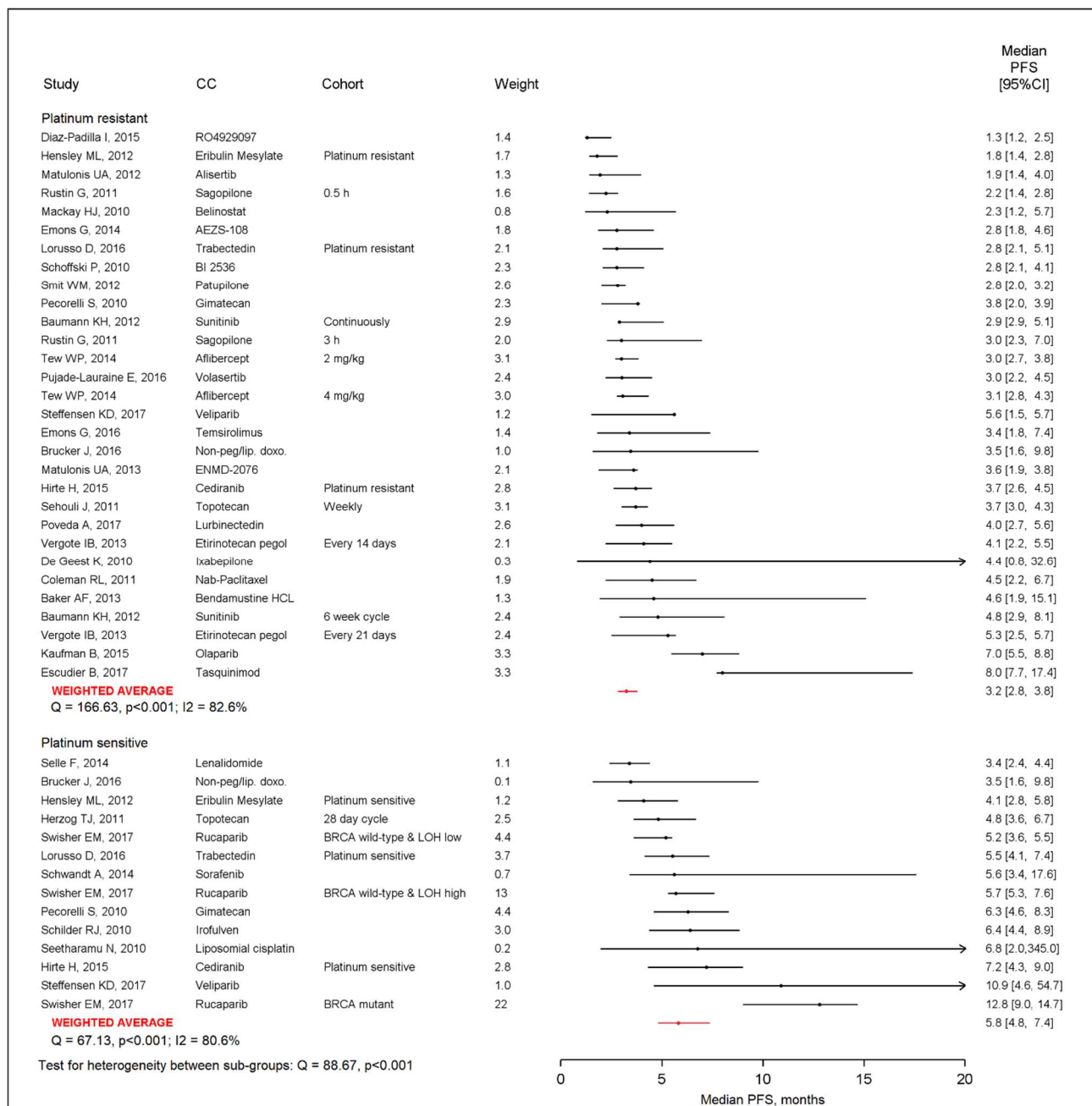


Figure 4.2.4 Median PFS distribution by platinum sensitivity

CC	ORR		Median PFS	
	N° of cohorts	Q (p-value), I ² (%)	N° of cohorts	Q (p-value), I ² (%)
Olaparib	7	6.19 (0.402), 3.0	7	32.00 (<0.001), 81.2
Sunitinib	5	3.65 (0.456), 0	5	6.71 (0.152), 40.4
Rucaparib	5	24.53 (<0.001), 83.7	3	37.64 (<0.001), 94.7
Topotecan	5	11.16 (0.025), 64.2	2	2.21 (0.137), 54.7
Sorafenib	3	2.73 (0.256), 26.7	3	16.11 (<0.001), 87.6
Cabozantinib	3	8.56 (0.014), 76.6	2	0.57 (0.450), 0
Veliparib	2	3.11 (0.078), 67.8	2	3.38 (0.066), 70.4
ENMD-2076	2	0.002 (0.962), 0	2	0.01 (0.913), 0
Temsirolimus	2	0.022 (0.883), 0	2	0.02 (0.883), 0

Table 4.2.2 Heterogeneity in different cohorts treated with the same CC

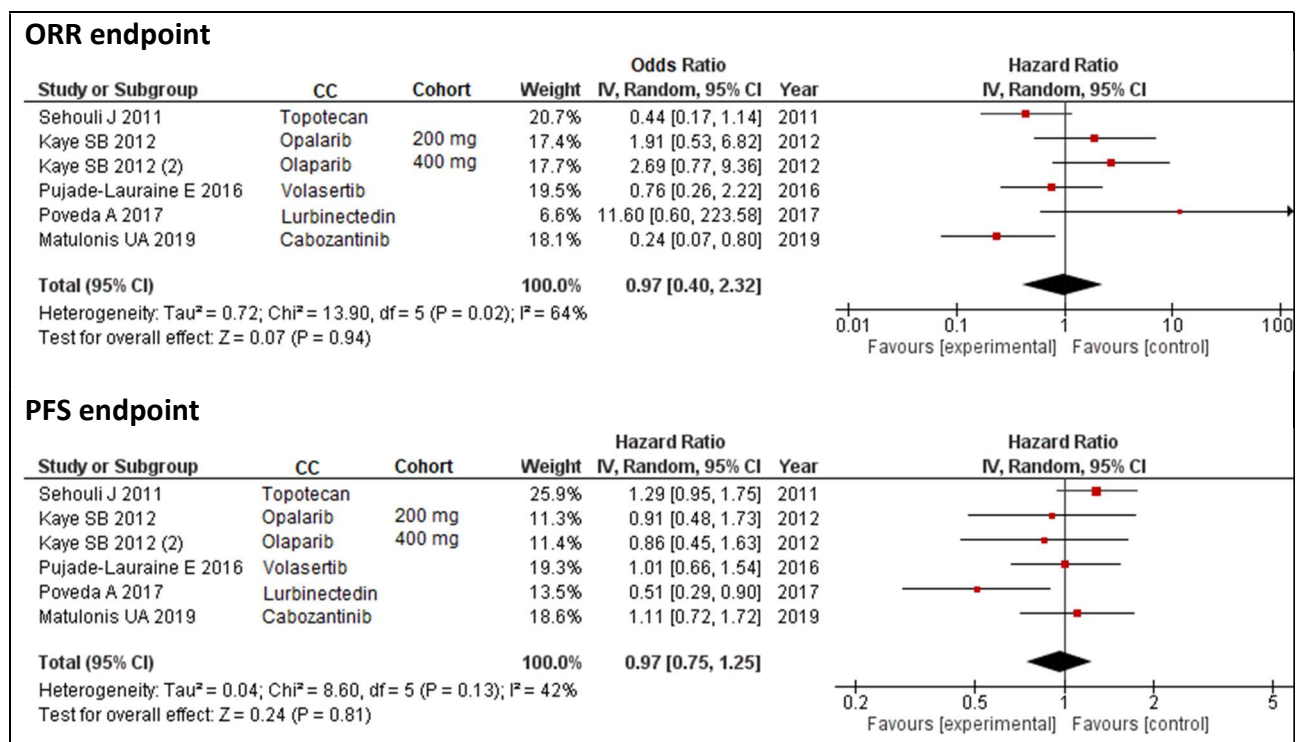


Figure 4.2.5 Heterogeneity of treatment effects in RCTs

Legend 4.2.5: Chi² values in the figure are the same as the Q statistic used previously. Forest plots were automatically generated using RevMan, version 5.3

4.3 Selection of preclinical *in vivo* antitumor activity studies

On 9th May 2019 a systematic search of the Medline and EMBASE databases was carried out to identify eligible preclinical *in vivo* antitumor activity studies and to assess their methodological quality. The systematic database search yielded 823 records. No additional records were identified by hand searching. After 58 duplicates were removed, a total of 765 articles remained. Due to poor reporting standards of preclinical *in vivo* antitumor activity studies, it was decided to screen articles based upon the full-text review. Six hundred ninety full-text articles were excluded on this basis, the major reasons being (1) no eligible CC was used in an *in vivo* EOC model, as monotherapy (n = 417, 60%), (2) no EOC model was used (n = 199, 29%), (3) the paper was requested to the Institute Library but it was not available (n = 43, 6%) and the paper was written in languages other than English (n = 28, 4%; refer to Figure 4.3.1). Seventy-five articles in total were included in this survey. They are listed in Appendix D.

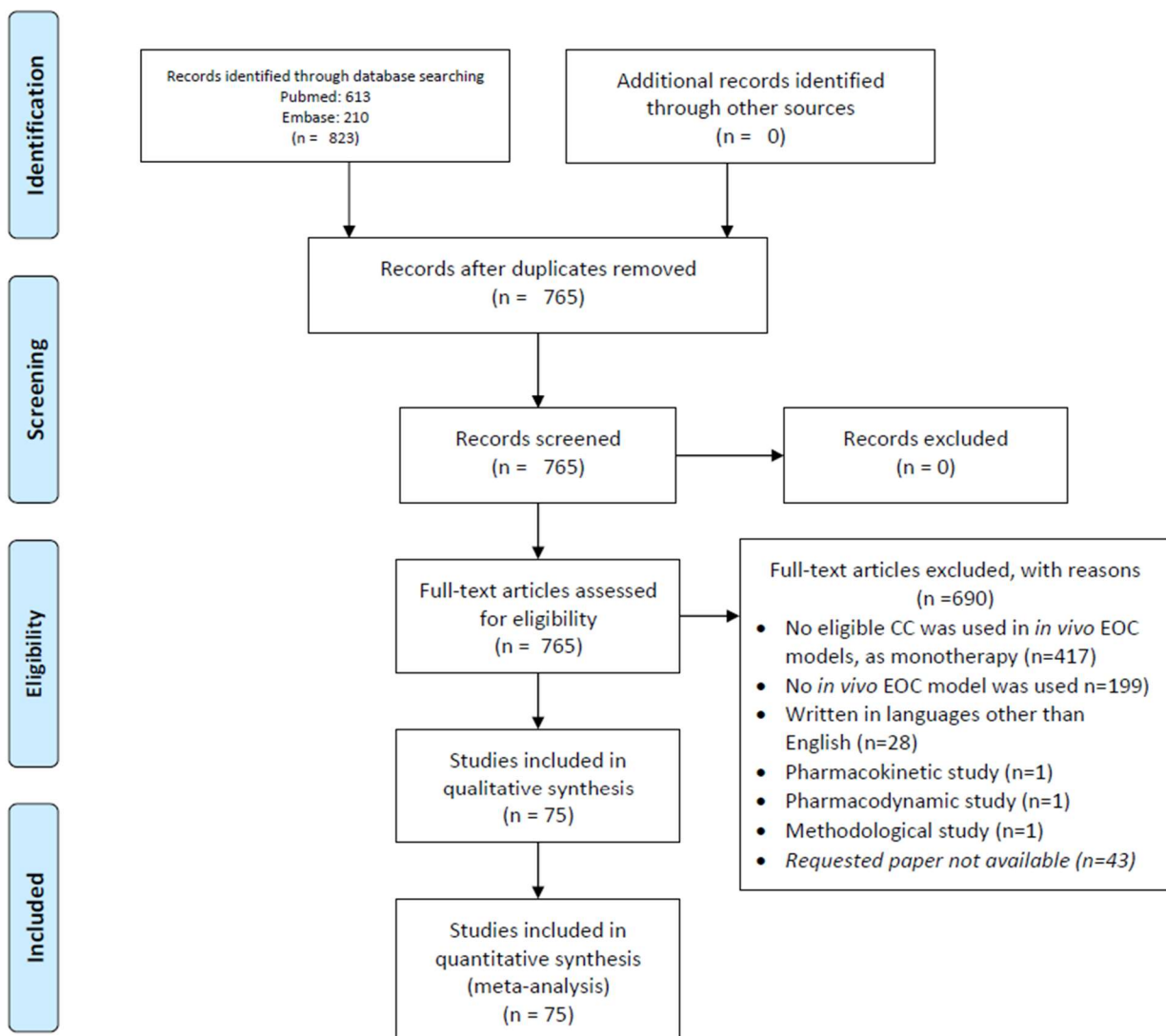


Figure 4.3.1 PRISMA flow diagram on selection of preclinical *in vivo* antitumor activity studies

Characteristics of eligible preclinical research articles are reported in Table 4.3.1. Fifty-three out of 75 (71%) eligible papers were published in the last ten years. One hundred twenty-one eligible *in vivo* experiments and 176 eligible *in vivo* EOC models were identified. Only 14 (19%) papers reported more than two eligible *in vivo* experiments. In 94 out of 121 (78%) eligible *in vivo* experiments, the CC was administered as monotherapy in a single eligible EOC model. The conflict of interest was declared in 40 (53%) papers. In 60 (80%) papers, the translation of preclinical results to the clinical setting was recommended or at least it was considered possible.

	<i>N</i>	%
Year of publication		
1998-2003	10	13
2004-2008	12	16
2009-2013	24	32
2014-2018	29	39
Number of eligible <i>in vivo</i> experiments, by eligible paper		
1	49	65
2	12	16
3	10	13
4	3	4
5	0	0
6	1	1
Number of eligible <i>in vivo</i> models, by eligible experiment		
1	94	78
2	20	17
3	1	1
4	4	3
11	1	1
12	1	1
Conflict of interests		
Declared		
Yes	40	53
Conflicts of interest are reported	13	32
None	27	68
No	35	47

	<i>N</i>	%
Recommended clinical translation		
Yes	32	43
Possible	28	37
No	15	20

Table 4.3.1 Major characteristics of eligible preclinical research papers

In Table 4.3.2, the list of eligible CCs, administered as monotherapy in at least one eligible EOC model, is reported. For 28 out of 52 (54%) eligible CCs, no eligible preclinical *in vivo* experiment was identified.

	<i>N</i> [§]	% ^{§§}
Topotecan	28	23
Olaparib	14	12
Cediranib	8	7
Gimatecan	8	7
Dasatinib	7	6
Nab-Paclitaxel	7	6
Pazopanib	7	6
Sunitinib	7	6
Vandetanib	6	5
Sorafenib	5	4
Irofulven	4	3
Eribulin Mesylate	4	3
Trabectedin	4	3
Belinostat	3	2
Aflibercept	3	2
Selumetinib	3	2
Perifosine	3	2
Alisertib	2	2
Rucaparib	2	2
Lapatinib	2	2
Etirinotecan pegol	2	2
Ixabepilone	1	1
Imatinib Mesylate	1	1
Non-pegylated liposomal doxorubicin	1	1

Table 4.3.2 List of CCs evaluated in preclinical *in vivo* experiments

Legend 4.3.2: [§] Number of preclinical *in vivo* experiments ^{§§} Does not sum to 100% as in 6 out of 121 (5%) preclinical *in vivo* experiments two eligible CCs were administered as monotherapy [topotecan and pazopanib (2); sorafenib and sunitinib (1); cediranib and olaparib (1); nab-paclitaxel and topotecan (1); gimatecan and topotecan (1)]

4.4 Tumor models used in preclinical *in vivo* antitumor activity studies

The characteristics of eligible *in vivo* tumor models are reported in Table 4.4.1. All models were immunocompromised. Unexpectedly and without explanation, 2 out of 140 (1.4%) tumor models were male mice (Alvero *et al.*, 2007; Nagengast *et al.*, 2011; refer to Appendix D). Poor reporting or methodological mistake could explain this data. One hundred thirty-nine out of 176 (79.0%) tumor models were CDX. The most common types of cells were A2780 and the derived cells (n = 37, 26.6%), and SKOV3 and the derived cells (n = 30, 21.6%). Athymic nude mice were the most commonly used strain (n = 98, 57.0%).

	N	%
Species		
Mouse	175	99
Rat	1	1
Age (weeks)		
Min-max	3-14	
Missing data	52	30
Sex		
Female	138	98.6
Male	2	1.4
Missing data	36	20.5
Weight (grams)		
Min-max	14-28	
Missing data	171	97
Immune status		
Immunocompromised	176	100
Immunocompetent	0	0
Model type		
Cell derived xenograft (CDX)	139	79.0
Patient derived xenograft (PDX)	37	21.0
CDX, type of cell		
A2780 and derived cells	37	26.6
SKOV3 and derived cells	30	21.6

	<i>N</i>	%
HEY and derived cells	13	9.4
IGROV-1	9	6.5
OVCAR-3	7	5.0
RMG-II	7	5.0
PEO1 and derived cells	4	2.9
RMG-I	4	2.9
ES-2	3	2.2
MA148	3	2.2
OVCAR-5 and derived cells	3	2.2
Caov-3	2	1.4
EFO-27	2	1.4
Pat-7	2	1.4
PEO4 and derived cells	2	1.4
RMUG-S and derived cells	2	1.4
TOV-21G	2	1.4
C13	1	0.7
HO-8910	1	0.7
OVTOKO	1	0.7
OV-2008	1	0.7
OV-90	1	0.7
OW1	1	0.7
<i>Tumor-initiating cells derived from A2780/HEY cell lines</i>	<i>1</i>	<i>0.7</i>
Strain		
Athymic nude	98	56.0
CD-1 nude	8	4.6
Swiss nude	19	10.9
BALB/c Nude	15	8.6
CB17 SCID	4	2.3
SCID-beige	4	2.3
NOD/Scid	8	4.6
NOD/SCID gamma	18	10.3

	<i>N</i>	%
RNU Nude (Rat species)	1	0.6
<i>Missing data</i>	<i>1</i>	<i>0.6</i>

Table 4.4.1 Major characteristics of eligible *in vivo* tumor models

4.5 Assessment of the quality of statistical design and analysis

Results of the assessment of the quality of statistical design and analysis are reported in Tables 4.5.1-4.5.8, by ad hoc checklist's sections. All preclinical *in vivo* experiments used mice, except one that used rats. Ninety-four out of 121 (77.7%) preclinical *in vivo* experiments used a single cancer model. All preclinical *in vivo* experiments were monocenter. Forty-five out of 120 (37.5%) preclinical *in vivo* experiments did not specify the method of treatment assignment and 71 out of 74 (95.9%) preclinical *in vivo* experiments did not specify the type of randomization. The technique of allocation concealment and the use of randomization for housing animals within the animal room and for assessing study outcomes were totally missing. As reported in Table 4.5.2, CC was the active arm in 64 out of 121 (52.9%) preclinical *in vivo* experiments. Twenty-three (19.0%) preclinical *in vivo* experiments used a dose-response design (i.e. ≥ 3 doses were administered including the placebo arm). Fifty-seven (47.1%) preclinical *in vivo* experiments used a factorial design but interaction between design factors was never formally tested. The technique of blocking was never used, neither in statistical design nor in statistical analysis. Sample size was justified in only 9 (7.4%) preclinical *in vivo* experiments. In these few cases, sample size was determined using a power analysis but the magnitude of the targeted biological signal was never justified. Moderate heterogeneity was observed in the number of animals enrolled by arm (IQR: 6-10; min-max: 3-29).

The primary outcome measure was tumor volume in 64 (52.9%) preclinical *in vivo* experiments. Tumor volume was measured in different ways (mono, bi and tridimensional ways). Lack of standardization was detected in tumor volume determination and in the definition of tumor response (i.e. PR/CR) and definition of cure of animal. In only 14 out of 106 (13.2%) cases the antitumor activity endpoint (i.e. the target tumor volume after which the rodent is sacrificed and hence tumor growth is not observed anymore) was clearly reported. Outcome assessor was blinded in only 5 (4.1%) cases.

Hypothesis testing, which is necessary to detect treatment effects, was used in 99 (81.8%) cases, while treatment effect's estimates were reported in only 61 (50.4%) cases. In the analysis of tumor growth curves, parametric methods were preferred over non-parametric methods but their statistical assumptions were questioned in only 9 out of 67 (13.4%) cases. On the other hand, non-parametric methods such as the log-rank test were preferred in survival analysis. The inefficient one-way ANOVA was preferred over the efficient linear mixed-effects models for repeated measures. The analysis and reporting of tumor growth curves were definitely poor. It was never possible to review the process, starting from treatment assignment to the rodents that were actually

analyzed. The number of animals on which the plotted point was based was never reported in the plot of tumor growth curves. Hence it was impossible to estimate the variability of the biological signal. In only 5 (4.1%) cases, the event that determined the end of follow-up was clearly reported for each rodent. The length of follow-up, a key element to debate the magnitude of biological signal, was specified in only 43 (35.5.%) cases. The statistician could be identified in 14 (11.6%) cases and the use of a statistical software was reported in 58 (47.9%) cases.

Item	Category	Statistics	
α1. Is the same experiment repeated more than once by a single lab?	Yes	N (%)	3 (2.5)
	No	N (%)	118 (97.5)
	N° of repetition	2	N (%) 3 (100)
	Reasons for repetition, specify	Poor statistical design; randomization was introduced in the repeated experiment	N (%) 1 (100)
	Unknown	N (%)	2 (66.6)
α2. N° of different species in which the EOC experiment has been repeated	1 (i.e. mouse)	N (%)	120 (99.2)
	2 (i.e. mouse and rat)	N (%)	1 (0.8)
α3. N° of different cancer models in which the EOC experiment has been repeated	1	N (%)	94 (77.7)
	2	N (%)	20 (16.5)
	3	N (%)	1 (0.8)
	4	N (%)	4 (3.3)
	11 §	N (%)	1 (0.8)
	12 §	N (%)	1 (0.8)
α4. N° of participating laboratories	Monolab	N (%)	121 (100)
	Multilab	N (%)	0

Table 4.5.1 Repetition and external validity

Legend 4.5.1: § Patient-derived xenograft models have been used

Item	Category	Statistics		
β1. Is an internal control group used?	Yes	N (%)	120 (99.2)	
	No	N (%)	1 (0.8)	
β1 a. If yes, are animals randomly allocated to treatment?	Yes	N (%)	74 (98.7)	
	No	N (%)	1 (1.3)	
	Unknown	N (%)	45 (37.5)	
	Alternative method to randomization	Matching	N (%)	1 (100)
β1 b. If randomization is used, is the randomization method stated?	Yes	N (%)	3 (4.1)	
	No	N (%)	71 (95.9)	
	Specify randomization method	Stratified randomization	N (%)	3 (100)
β1 c. If randomization is used, is allocation concealment employed?	Unknown	N (%)	74 (100)	
β1 d. Are the animals randomly housed within the animal room?	Unknown	N (%)	121 (100)	
β1 e. Are there equal numbers per treatment group?	Yes	N (%)	82 (75.2)	
	No	N (%)	27 (24.8)	
	Unknown	N (%)	12 (9.9)	
β1 f. If not, is this justified?	No	N (%)	27 (100)	
β2. Is the experimental unit clearly identified?	Yes	N (%)	121 (100)	
	No	N (%)	0	
	Specify experimental unit	Single animal	N (%)	121 (100)
	Suspicious case of pseudo-replication	Drug was dissolved in cage’s water	N (%)	1 (0.8)
β3. Role of CC monotherapy arm	Active	N (%)	64 (52.9)	
	Only control arm	N (%)	57 (47.1)	

Table 4.5.2 Internal validity

Item	Category	Statistics	
γ1. Is an active control group used?	Yes	N (%)	83 (68.6)
	No	N (%)	38 (31.4)
γ2. Number of design factors	1	N (%)	64 (52.9)
	2	N (%)	53 (43.8)
	3	N (%)	4 (3.3)
γ3. Number of treatment (design) groups		N	121
		Median	4
		IQR	3-5
		Min-max	2-12
γ4. Is a factorial (complete or incomplete) design used?	Yes	N (%)	58 (47.9)
	No	N (%)	63 (52.1)
γ5. Is dose-response (i.e. ≥ 3 doses) evaluated?	Yes	N (%)	23 (19.0)
	No	N (%)	98 (81.0)
γ6. Is blocking used?	Yes	N (%)	0
	No	N (%)	121 (100)
γ7. Type of experiment	Between units	N (%)	121 (100)
	Within units	N (%)	0
	Both	N (%)	0

Table 4.5.3 Statistical design

Item	Category	Statistics	
δ1. Is the sample size justified?	Yes	N (%)	9 (7.4)
	No	N (%)	112 (92.6)
δ1 a. If yes, specify method	Common sense	N (%)	0
	Power analysis	N (%)	9 (100)
	Resource equation	N (%)	0
	Other method to determine sample size	N (%)	0
δ2. Total number of enrolled animals per arm [§]	N	N	116
	Median	Median	9
		IQR	6-10
		Min-max	3-29
	Unknown	N (%)	5 (4.1)

Table 4.5.4 Sample size

Legend 4.5.4: [§] Both methods and results sections were considered to calculate the total number of enrolled animals per arm

Item	Category	Statistics			
ε1. Can the primary outcome measure of antitumor activity clearly be identified? §	Yes	N (%)	121 (100)		
	No	N (%)	0		
ε2a. Specify primary outcome measure	Tumor volume	N (%)	64 (52.9)		
	Tumor weight	N (%)	30 (24.8)		
	Overall survival	N (%)	15 (12.4)		
	Photon counts per area	N (%)	8 (6.6)		
	Volume of ascites	N (%)	2 (1.7)		
	Number of metastases	N (%)	1 (0.8)		
	Number of cured animals	N (%)	1 (0.8)		
ε2b. Specify primary or secondary outcome measures	Tumor volume	N (%)	68 (56.2)		
	Tumor weight	N (%)	37 (30.6)		
	Overall survival	N (%)	20 (16.5)		
	Number of tumor nodules	N (%)	18 (14.9)		
	Time to reach a target volume	N (%)	14 (11.6)		
	Photon counts per area	N (%)	9 (7.4)		
	Volume of ascites	N (%)	5 (4.1)		
	Number of metastases	N (%)	2 (1.7)		
	Number of cured animals	N (%)	2 (1.7)		
	Tumor burden §§	N (%)	1 (0.8)		
	Definition of tumor volume §§§	Monodimensional	(length) ³ x (π/6)	N (%)	1 (0.02)
			(length) ³ x (π/6)	N (%)	1 (0.02)
	Bidimensional	(length x width ²) / 2	N (%)	24 (48.0)	
		(length x width ²) x (π/6)	N (%)	7 (0.14)	
		(length x width)/2	N (%)	2 (0.04)	
		(length) ^{3/2} x width ² x (π/6)	N (%)	1 (0.02)	
		(length x width) ^(3/2) x (π/6)	N (%)	1 (0.02)	

Definition of cured animals Measures of treatment effect	Tridimensional	(length x width) x (π/3)	N (%)	1 (0.02)	
		[(length + width)/4] ³ · (0.67 · π)	N (%)	1 (0.02)	
		(length x width x height) / 2	N (%)	6	
		(length x width x height) x (π/6)	N (%)	4 (0.1)	
		(length x width x height) x (4 · π/3)	N (%)	2 (0.04)	
	Missing data		N (%)	20 (28.2)	
	Animals without tumor at the end of the study		N (%)	1 (50.0)	
		Recovery of the initial weight	N (%)	1 (50.0)	
	Tumor volume	T/C	N (%)	30 (44.1)	
		CR/PR ^{§§§§}	N (%)	14 (20.6)	
		% change respect to treatment start	N (%)	4 (5.9)	
		Time to reach a target volume ^{§§§§§}	Log Cell Kill (LCK)	N (%)	4 (28.6)
			Tumor Growth Delay (TGD)	N (%)	3 (21.4)
ε3. Is the antitumor activity endpoint (i.e. event) clearly defined?	Overall survival as primary outcome measure	Yes	N (%)	13 (86.7)	
		No	N (%)	2 (13.3)	
	Tumor volume and other primary outcome measures	Yes	N (%)	14 (13.2)	
		No	N (%)	92 (86.8)	
ε4. Are the competing events clearly defined?		Yes	N (%)	3 (2.5)	
		No	N (%)	118 (97.5)	
ε5. Outcome assessment					
ε5 a. Are the caregivers and/or investigators blinded?		Yes	N (%)	1 (0.8)	
		No	N (%)	120 (99.2)	
ε5 b. Are animals selected at random for outcome assessment?		Unknown	N (%)	121 (100)	
ε5 c. Is the outcome assessor blinded?		Yes	N (%)	5 (4.1)	
		No	N (%)	116 (95.9)	

Table 4.5.5 Outcomes and their assessment

Legend 4.5.5: [§] Only one eligible paper defined clearly the primary outcome (Decio *et al.*, 2015; refer to Appendix D). In all other cases, assessor's judgement deduced which was the primary outcome based on the report and the biological relevance. Reporting bias about primary/secondary outcomes could be present although it was impossible to detect it ^{§§} An arbitrary score was used to estimate tumor dissemination ^{§§§} 71 preclinical *in vivo* experiments used as outcome measure the tumor volume, measured by caliper, or the time to reach a target volume ^{§§§§} CR/PR: Complete Response/Partial Response; CR and PR were defined differently by eligible papers ^{§§§§§} 14 preclinical *in vivo* experiments used the time to reach a target volume as secondary outcome. LCK is defined as $[(T-C) / 3.32] \times DT$, where T and C are the mean times (days) required for treated (T) and control (C) tumors, respectively, to reach the target volume, and DT is the doubling time of control tumors. TGD is defined as T-C, where T and C are the mean times (days) required for treated (T) and control (C) tumors, respectively, to reach the target volume

Item	Category	Statistics	
ζ1. Are inferential methods used to demonstrate antitumor activity?	Yes	N (%)	106 (87.6)
	No	N (%)	15 (12.4)
ζ1 a. Hypothesis test	Yes	N (%)	99 (93.4)
	No	N (%)	7 (6.6)
	Specify hypothesis test	One-way parametric ANOVA	N (%) 56 (56.6)
		Log-rank	N (%) 20 (20.2)
		One-way non-parametric ANOVA	N (%) 13 (13.1)
		Two-way parametric ANOVA for repeated measures	N (%) 11 (11.1)
		Other hypothesis tests [§]	N (%) 5 (5.1)
		<i>Not specified hypothesis test</i>	N (%) 2 (2.0)
ζ1 b. Estimation	Yes	N (%)	61 (57.5)
	No	N (%)	45 (42.5)
	Specify estimator	Mean plus SE or 95%CI	N (%) 60 (98.4)
		Hazard Ratio plus 95%CI	N (%) 1 (1.6)
ζ2. Are descriptive methods used to demonstrate antitumor activity?	Yes	N (%)	109 (90.1)
	No	N (%)	12 (9.9)
	Specify methods ^{§§}	Tumor growth curve	N (%) 54 (49.5)
		Mean plus STD	N (%) 23 (21.1)
		Kaplan-Meier	N (%) 17 (15.6)
ζ3. Are methods for correction of multiple comparison used?	Yes	N (%)	20 (22.7)
	No	N (%)	68 (77.3)
		<i>Not applicable</i>	N (%) 33 (27.3)
	Specify methods	Post-hoc tests ^{§§§}	N (%) 19 (95.0)
		Holm's procedure	N (%) 1 (5.0)

4. Are statistical assumptions used to analyze tumor growth data justified?	Yes	N (%)	9 (13.4)
	No	N (%)	58 (86.6)
	<i>Not applicable</i>	N (%)	54 (44.6)
	Specify methods	N (%)	7 (77.8%)
	Normality tests ^{§§§§}		
	Log-transformation		2 (22.2%)

Table 4.5.6 Statistical analysis

Legend 4.5.6: [§] Wilcoxon matched-pairs signed test (3 preclinical *in vivo* experiments); Fisher's exact test (1 preclinical *in vivo* experiments), Q test for heterogeneity (1 preclinical *in vivo* experiment) ^{§§} Only descriptive methods used by a number of preclinical *in vivo* experiments greater or equal to 5 are reported ^{§§§} Tukey (9 preclinical *in vivo* experiments); Bonferroni (5 preclinical *in vivo* experiments); Dunnett (5 preclinical *in vivo* experiments) ^{§§§§} Kolmogorov-Smirnov test (6 preclinical *in vivo* experiments), Shapiro-Wilk test (1 preclinical *in vivo* experiment)

Item	Category	Statistics	
η1. Number of animals assigned to each treatment arm (results section)	Exact	N (%)	67 (55.4)
	Estimate	N (%)	16 (13.2)
	Unknown	N (%)	38 (31.4)
η2. Are there animals assigned to each treatment arm and excluded from statistical analysis? Specify reasons for exclusion	Unknown	N (%)	121 (100)
		N (%)	-
η3. Are there animals at risk progressively reported in the plot of tumor growth curves?	Yes	N (%)	0
	No	N (%)	54 (100)
	Not applicable [§]	N (%)	67 (55.4)
η4. Is the number of animals with right-censored data clearly reported?	Yes	N (%)	5 (4.1)
	No	N (%)	116 (95.9)
η5. For each animal, is it clearly reported which event determined the end of follow-up?	Yes	N (%)	5 (4.1)
	No	N (%)	116 (95.9)
η6. Is the length of follow-up clearly defined? Specify definition	Yes	N (%)	43 (35.5)
	No	N (%)	78 (64.5)
	Based on treatment schedule	N (%)	33 (76.7)
	Control group reached critical conditions	N (%)	7 (16.3)
	The target volume was reached	N (%)	2 (4.7)
	Historical growth rate of the tumor model	N (%)	1 (2.3)

Table 4.5.7 Attrition bias about tumor growth curves

Legend 4.5.7: [§] Tumor growth curves were reported in 54 preclinical *in vivo* experiments (refer to Table 4.5.6)

Item	Category	Statistics	
01. State any important concerns about statistical design and analysis not covered by other sections in the checklist [§]	Interaction was not formally evaluated in a factorial design	N/Tot (%)	56/57 (98)
	One-way parametric/non-parametric ANOVA was repeated for all time points	N/Tot (%)	9/54 ^{§§} (17)
	Mice were grafted bilaterally. Statistical analysis was not explained	N/Tot (%)	5/6 (83)
	Different number of animals reported in methods and results sections	N/Tot (%)	5/116 (4)
	In survival analysis, lead time bias was detected using the day of tumor transplant, instead of the day of randomization, as the starting day	N/Tot (%)	3/20 ^{§§§} (15)
02. Was any author a member of a department of statistics or epidemiology?	Yes	N (%)	14 (11.6)
	No	N (%)	107 (88.4)
03. Was mentioned the use of a statistical software for data analysis?	Yes	N (%)	58 (47.9)
	No	N (%)	63 (52.1)
	Specify statistical softwares ^{§§§§}	GraphPad Software, La Jolla, CA, USA	N (%) 38 (65.5)
		SPSS Inc, Chicago, IL	N (%) 14 (24.1)
		Excel	4 (6.9)
		STATA	3 (5.2)
		SAS	N (%) 2 (3.4)
		JMP	N (%) 1 (1.7)
		Python	N (%) 1 (1.7)
		R	N (%) 1 (1.7)

Table 4.5.8 Miscellanea

Legend 4.5.8: [§] Only statistical pitfalls regarding a number of preclinical *in vivo* experiments greater or equal to 3 are reported ^{§§} Tumor growth curves were reported in 54 preclinical *in vivo* experiments ^{§§§} Overall survival was the primary or secondary outcome in 20 preclinical *in vivo* experiments ^{§§§§} Two statistical softwares were used in 6 preclinical *in vivo* experiments

4.6 Correlation between preclinical quality and phase 2 activity

There were no preclinical *in vivo* experiment eligible to be classified with good methodological quality. Even if two of the items defining a 'good' *in vivo* experiment are not considered, no preclinical *in vivo* experiment could be classified with good methodological quality. Different preclinical *in vivo* experiments had 'good' scores on different aspects. But different aspects could not be used confidently to rank the quality of preclinical *in vivo* experiments.

In conclusion, due to poor methodology applied, it was impossible to correlate quality of preclinical *in vivo* experiments with clinical activity.

Chapter 5

Improving statistics of *in vivo* tumor growth curves

5.1 Concepts

This section reviews the concepts of hypothesis testing and estimation. Readers who are familiar with these concepts can skip to the Section 5.2.

5.1.1 Testing statistical hypotheses

The researcher is primarily interested in detecting a treatment effect. The definition of 'no effect' is as follows: each animal exhibits a response (i.e. outcome) that is observed some time after treatment. To say that the treatment has no effect on this response is to say that each animal would exhibit the same value of the response whether assigned to active or control arm. If changing the treatment assigned to an animal changed that animal's response, then certainly the treatment has at least some effect. Because it is usually impossible to observe animal's response under both conditions (i.e. each animal is treated according to the active and control arm), animals are randomized to the active and control arm and their responses are collected. A test statistic (e.g. mean difference or the Mann-Whitney statistic) is then calculated from the data. The primary usefulness of the randomization procedure is that the probability distribution of the test statistic when the null hypothesis (i.e. no treatment effect) is true is known. Based on this probability distribution, a 'p-value' is assigned to the test statistic. The p-value is the probability of obtaining a test statistic as extreme or more extreme when the null hypothesis is true. If the p-value is less than a sufficiently small value, called alpha (i.e. the significance level of the test; for example, 0.05), the investigator can 'reject the null hypothesis' and conclude that the treatment really does have an effect. A rejection of the null hypothesis when it is true is an error of inference, known as a type I error or 'false positive' result. An alpha (α) of 0.05 means that the investigator is willing to accept a type I error 5% of the time if the null hypothesis (i.e. no treatment effect) is actually true. If the treatment has no effect on the chosen outcome and the researcher replicates the same experiment 100 times, we can expect that the significance level of the test is reached in 5 replications. The significance level of the test should be determined in advance, in the same manner that the rules of a card game are defined before the gamblers play with dealers.

The p-value does not provide any information about how large or important the treatment effect is. Although the p-value does not provide any information about the treatment effect's estimate, it is strictly related to the accuracy of the treatment effect's estimate. Increasing sample

size, even tiny and unimportant effects can be discovered and accuracy of treatment effect's estimate is improved.

There is a second type of error of inference. A type II error occurs when one fails to discover a treatment effect when the null hypothesis is false (i.e. 'false negative' result). The probability of type II error is denoted as β . The complement of type II error, the power (i.e. $1 - \beta$), is the probability that the null hypothesis will be rejected correctly given that there is a treatment effect. The power is a function of the treatment effect and sample size, as shown in Figure 5.1.1.1.

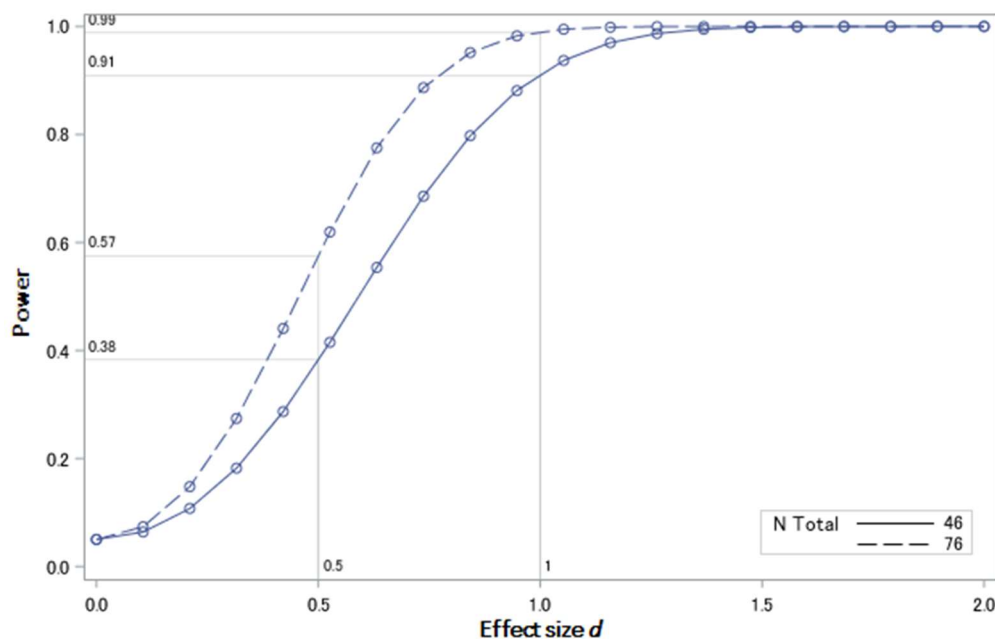


Figure 5.1.1.1 Power as a function of treatment effect, expressed as effect size d (i.e. difference between means divided by standard deviation) and total sample sizes (i.e. "N Total") in a 2-tailed t test with 2 independent groups, allocation ratio 1:1 and $\alpha = 0.05$

An *in vivo* experiment is underpowered if the power is inadequate (e.g. power = 0.50 or 0.60) to detect a valuable treatment effect. As consequence of underpowered *in vivo* experiments, animals could be wasted because the probability of type II error is high. An *in vivo* experiment is overpowered if power is too large. As consequence of overpowered *in vivo* experiments, trivial effects could be detected and too many animals are used. For example, as shown in Figure 5.1.1.1, if an effect of 1 standard deviation (STD) is the smallest interesting effect, designing the experiment to detect this effect with high power, such as 99%, could waste animals (total sample size is 76 animals), because the test continues to detect trivial effects, such as 0.5, with a high frequency (57%). Setting a power of 90% for the smallest interesting effect, with a total sample size of 46 animals, causes the detection of trivial effects to decline more steeply.

5.1.2 Point and interval estimates

Once the treatment effect has been detected, a researcher is usually interested to estimate it. To solve this problem, the observations collected from an *in vivo* experiment are postulated to be the values taken by random variables which are assumed to follow a joint probability distribution. Eisenhart (1964) attributes the crucial step of considering observed data as generated by random variables to Simpson (1755).

The probabilistic model that generates *in vivo* experiment's observations is usually indexed by a set of parameters. The aim of the analysis is then to specify a plausible value for each parameter (this is the problem of point estimation), or at least to determine a subset of plausible values for each parameter (estimation by confidence intervals or credible intervals).

For example, the probabilistic model of the one-way ANOVA is the following:

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}, i=1,...,t; j=1,...,r_i \quad 5.1.2.1$$

where Y_{ij} is the outcome and the parameters μ and τ_i are the overall mean and the effect of treatment i , respectively. ε_{ij} is the error (i.e. the deviation of the particular unit j from the average of the set of units assigned to treatment i), often taken to be independent, identically and normally distributed with mean equal to zero. Casella's book (2008) brilliantly discusses this model. If the experimental data are generated by the ANOVA model, the researcher is interested to estimate the parameters $\mu + \tau_i$, with $i=1,...,t$.

An estimator of a parameter is said 'unbiased' if its expected value is equal to the true value of the parameter. For example, from 5.1.2.1

$$E \{ (1 / r_i) \cdot [\sum_j (\mu + \tau_i + \varepsilon_{ij})] \} = \mu + \tau_i \quad 5.1.2.2$$

showing that the parameter $\mu + \tau_i$ is estimable and that the sample mean is an unbiased estimator.

Confidence intervals and credible intervals are common measures used to specify subsets of plausible values for a probabilistic model's parameter. Confidence and credible intervals are estimates of an interval that may contain the true value of the parameter. The interval is generally defined by its lower and upper bounds. A confidence or credible interval is usually expressed as a percentage (the most frequently quoted percentages are 90%, 95%, and 99%). The percentage is called the confidence and credible level, respectively. The confidence interval is constructed in the following way: parameter values that when postulated as the null hypothesis produce a p-value greater than $1 - \text{percentage} / 100$ (e.g., $1 - 90/100 = 0.10$, $1 - 95/100 = 0.05$ and $1 - 99/100 = 0.01$), will typically define a range of values that would be considered more compatible with the data than values outside the range - if the probabilistic model generating *in vivo* experiment's observations is

correct. This range of values corresponds to a percentage (e.g. 90%, 95% and 99%) confidence interval, and provides a convenient way of summarizing the results of hypothesis tests for many parameter values. A credible interval is constructed in a way so that an unobserved parameter value falls within this interval with a particular probability. In bayesian statistics, the parameter is itself a random variable with a known probability distribution. Hence, credible intervals are commonly used to summarize the posterior probability distribution.

5.2 Non-parametric Two-Sample Tests

5.2.1 Definition of statistical tests

When the tumors have reached a designated target volume, tumor-bearing animals are randomly assigned to control and experimental arms. Inactive vehicle or antitumor therapies are delivered to them, respectively. Tumor volumes are then measured with a caliper at randomization ($0 = t_0$) and periodically (e.g. twice a week) at common time points t_r , $r=1,...,K$, until the tumors reach a predetermined target volume.

For some animals, tumor volume does not reach the target size because competing events occurs. For instance, death due to toxic effects of antitumor therapies or aggressive behaviour among animals occurs, a humane endpoint (i.e. the earliest indicator used to avoid or stop the distress, discomfort, or potential pain and suffering) is reached and animal sacrifice is ethically necessary, or a planned time to complete the experiment is reached. Tumor-bearing animals, for whom the predetermined target volume have not been observed, are said to provide right-censored data.

Formally, the null hypothesis of interest is that tumor volumes are equally distributed in control and treatment arms, for each time t_r , $r=0,...,K$. Notice that only animals, whose tumor does not reach the predetermined target volume, participate in the assessment of tumor volume distribution at time t_r , $r=1,...,K$. To simplify statistical considerations, let us suppose that common time points are equidistant [i.e. $t_r - t_0 = r \times (t_1 - t_0)$, $r=1,...,K$]. Let $v_i(t_r)$ be the tumor volume at time t_r of the i th animal. Each couple of subsequent and nonmissing tumor volume data $v_i(t_r)$, $v_i(t_{r+1})$, $r=0,...,K-1$ are summarized as $g_{r,i}$, using a slope function that increases monotonically with the distance between $v_i(t_r)$ and $v_i(t_{r+1})$. For instance, $g_{r,i}$ could be defined as $\{ v_i(t_{r+1}) - v_i(t_r) \} / (t_{r+1} - t_r)$, $\{ v_i(t_{r+1}) - v_i(t_r) \} / (t_{r+1} - t_r)^2$, $\{ \log(v_i(t_{r+1})) - \log(v_i(t_r)) \} / (t_{r+1} - t_r)$ or $\{ 1 + v_i(t_{r+1}) \} / \{ 1 + v_i(t_r) \}$ using respectively a linear, quadratic, exponential or ratio function. For each interval $[t_r, t_{r+1}]$, $r=0,...,K-1$, the Mann-Whitney W_r statistic, its expected value (E_r) and variance [$\text{var}(W_r)$] could be calculated on the samples of observations $g_{r,i}$ computed from the control and experimental arms. Let $X_{r,1},...,X_{r,m(r)}$

be the sample of slopes $g_{r,i}$ computed from the experimental arm, identically and independently distributed with probability distribution function $T_r(x)$, for all $x \in (-\infty, +\infty)$, at time t_r , $r=0, \dots, K-1$. Let $Y_{r,1}, \dots, Y_{r,n(r)}$ be the sample of slopes $g_{r,j}$ computed from the control arm, identically and independently distributed with probability distribution function $C_r(x)$, for all $x \in (-\infty, +\infty)$, at time t_r , $r=0, \dots, K-1$. Consider all possible pairs of slopes $(X_{r,i}, Y_{r,j})$ consisting of one slope $g_{r,i}$ computed from experimental arm and one slope $g_{r,j}$ computed from control arm, at time t_r , $r=0, \dots, K-1$. The Mann-Whitney W_r statistic at time t_r , $r=0, \dots, K-1$, is defined as

$$[\text{number of pairs } (X_{r,i}, Y_{r,j}) \text{ with } X_{r,i} < Y_{r,j}] + \frac{1}{2} \cdot [\text{number of pairs } (X_{r,i}, Y_{r,j}) \text{ with } X_{r,i} = Y_{r,j}] \quad 5.2.1.1$$

As consequence of the central limit theorem, the null distribution of the W_r statistic is asymptotically normal. Assuming no ties in tumor volume data,

$$E_r = m_r \cdot n_r / 2 \quad 5.2.1.2$$

$$\text{var}(W_r) = m_r \cdot n_r \cdot (m_r + n_r + 1) / 12 \quad 5.2.1.3$$

where m_r and n_r are respectively the number of animals in the experimental and control arms at time t_r , $r=0, \dots, K-1$. In presence of tied slopes, E_r is the same while $\text{var}(W_r)$ changes in the following more complex formula:

$$\{ m_r \cdot n_r \cdot (m_r + n_r + 1) / 12 \} - \{ (m_r \cdot n_r \cdot (\sum_j (d_j^3 - d_j))) / (12 \cdot (m_r + n_r) \cdot (m_r + n_r - 1)) \} \quad 5.2.1.4$$

for $j=1, \dots, e$, where e is the number of distinct values taken by the $m_r + n_r$ observations and d_j is the number of observations equal to the j th distinct value. The first term of the formula is just the variance of W_r in absence of tied data; the second gives the correction for ties. When no ties are present, all the d_j are equal to 1, and the correction term is zero as it should be. Refer to the Lehmann's book (2006) for proofs of previous formulas.

Once the previous statistics have been computed in each interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$, they can be combined into an overall test. The overall standardized test statistic can be defined as

$$\{ \sum_r (W_r - E_r) \}^2 / \text{var}(\sum_r W_r) \quad 5.2.1.5$$

for $r=0, \dots, K-1$. This statistic is approximately distributed as chi-square with one degree of freedom. It follows from the subsequent result in probability theory: a random variable, which is the sum of correlated normal random variables, is also distributed according to a normal distribution, with its mean being the sum of the means, and variance being the sum of the elements of the covariance matrix.

A one-sided test could be performed using the following statistic:

$$\sum_r (W_r - E_r) / (\text{var}(\sum_r W_r))^{1/2} \quad 5.2.1.6$$

$r=0, \dots, K-1$. The previous statistic has an approximate standard normal distribution.

5.2.2 Censoring and missing data

The family of statistical tests proposed in Section 5.2.1 are valid (i.e. statistical test distribution under null hypothesis is exactly known) in case of non-informative right-censoring. Formally, the condition of non-informative right-censoring observations in our *in vivo* tumor efficacy experiments could be defined as follows: at time t_r , $r=0, \dots, K-1$, let $m_r + n_r$ be the number of animals with tumor volume measurement, where m_r and n_r are respectively the number of animals in the experimental and control arms. At time t_{r+1} , $r=0, \dots, K-1$, suppose that c out of $m_r + n_r$ animals lack a tumor volume measurement due to competing events. These c animals will provide non-informative right-censoring observations if they are randomly selected from the $m_r + n_r$ animals.

An equivalent definition of non-informative right-censoring observations is the following: let $T_r(x)$ and $C_r(x)$ for all $x \in (-\infty, +\infty)$, at time t_r , $r=0, \dots, K-1$, be the probability distribution functions generating the sample of slopes computed respectively from the experimental and control arms. To the i th tumor-bearing animal, a couple of data (t_i, δ_i) could be assigned, where t_i is the observed time since start of inactive or active treatment and δ_i is an indicator of failure, assuming values $\delta_i = 1$ if the predetermined target volume has been reached and $\delta_i = 0$ if it has not been reached. Then survival data of tumor-bearing animals could be represented by the following U , V and T random variables: U is the survival time to reach the predetermined target volume, which we cannot always observe, and V is the potential censoring time. The observed time T is:

$T = \min(U, V)$ and $\delta = 1$ if $U \leq V$ or $\delta = 0$ if $U > V$. $T = U$ only when the observation is not right-censored.

In our *in vivo* tumor efficacy experiments, right-censoring is non-informative if and only if both condition holds:

- the random variables $T_r(x)$ and V are independent, for all times t_r , $r=0, \dots, K-1$ 5.2.2.1
- the random variables $C_r(x)$ and V are independent, for all times t_r , $r=0, \dots, K-1$ 5.2.2.2

However, random variables U and V could be dependent without contradicting the condition of non-informative right-censoring.

In practical terms, non-informative right-censoring means that the tumor growth experience of censored animals can be fairly estimated by using the data on the uncensored animals available at the time points following right-censoring. This is true also in case of non-informative missing data at common time points $t_0 < \dots < t_K$, until the censoring date. In case of non-informative right-censoring or missing data at time t_r , $r=0, \dots, K$, slope estimates are unbiased and statistical tests proposed in Section 5.2.1 are valid.

For specific competing events, such as administrative censoring, when V is the planned time to complete the experiment, the condition of non-informative right-censoring could be fairly assumed. For other competing events, this condition is doubtful or definitively questionable. For instance, suppose that a humane endpoint at time t_r , $r=1,...,K$, has been reached and a particular animal in the experimental arm must be sacrificed. One could argue that tumor growth slope of this animal could be very large in the time interval $[t_r, t_{r+1}]$ and so, the independence between random variables $T_r(x)$ and V is definitively questioned.

If our interest is to test treatment effect in absence of certain biological conditions (e.g., the animal is moribund) the family of statistical tests proposed in Section 5.2.1 are again valid. Clearly, what permits us to apply the same statistical procedure is that our aim is to detect treatment effect conditional on specific biological conditions. Notice that measuring tumor volumes below the predetermined target volume is the first biological constrain that we apply to our *in vivo* tumor efficacy experiments. In conclusion, our family of statistical tests is valid not only in case of non-informative right-censoring observations but also when the experimental models are constrained by more specific condition (e.g. tumor volume below the predetermined target volume, *in vivo* models not in moribund status).

Conversely, if our interest is to test treatment effect incorporating competing events that cause informative right-censoring observations, the family of statistical tests proposed in Section 5.2.1 does not hold. The useful null hypothesis to be investigated is that the cumulative probability functions $F(t_r)$, $r=1,...,K$, are the same in control and treatment arms. For each $r=1,...,K$, $F(t_r)$ is the probability distribution function generating the sample of tumor growth curves defined in the time interval $[t_0-t_r]$, in presence of competing events. Animals affected by informative right-censoring at time t_{r^*} , $r^*=2,...,K$, contribute to the definition of each function $F(t_r)$, $r < r^*$. The study of this function is not the subject of this project. In survival analysis, we refer to Kalbfleisch and Prentice who suggested the study of an analogue cumulative probability function, the cumulative incidence function (Kalbfleisch *et al.*, 1980). This approach, that accounted for the presence of competing events, led to the formulation of the Gray's test (Gray, 1988) and of the Fine and Gray regression model (Fine *et al.*, 1999).

5.2.3 Weighted non-parametric Two-Sample Tests

A number of weighted tests can be generated from the basic tests defined in Section 5.2.1. These are generally defined as

$$\{ \sum_r [w_r \cdot (W_r - E_r)] \}^2 / [\text{var}(\sum_r w_r \cdot W_r)] \quad 5.2.3.1$$

for $r=0, \dots, K-1$, where w_r are positive constants. The weights w_r can vary as time t_r , $r=0, \dots, K-1$, changes. The statistic reported in Formula 5.2.3.1 is asymptotically distributed according to the chi-square density function with 1 d.f. for the two group comparisons discussed here, under the null hypothesis that tumor volumes are equally distributed in control and treatment arms.

If we set $w_r = 1$ or broadly $w_r = \text{constant}$, for $r=0, \dots, K-1$, then this assigns equal weight to each part of the growth curve. In this situation, Equation 5.2.3.1 becomes that of the unweighted test of Equation 5.2.1.5. Instead, if we set $w_r \neq w_{r+1}$ for at least one r , $r=0, \dots, K-1$, we are implicitly stating that differences between certain parts of the growth curves being compared are of greater interest than others. For example, if it is anticipated that the active treatment reduces the tumor growth rate at the start of treatment delivery but thereafter holds no particular advantage, then higher weight may be assigned to the earlier parts of the growth curves.

In survival analysis, weighted Mantel-Haenzel (M-H, log-rank) tests can be generated from the basic formula of the unweighted M-H test according to the same Equation 5.2.3.1. Expected value and variance of the Mann-Whitney W_r statistic are substituted respectively by expected value and variance of the M-H statistic. Two interesting types of weighted M-H tests have been proposed, according to the following values given to the weights w_r , $r=0, \dots, K-1$:

$$w_r = m_r + n_r \quad (\text{Gehan test; Gehan, 1965}) \quad 5.2.3.2$$

$$w_r = \sqrt{m_r + n_r} \quad (\text{Tarone and Ware test; Tarone et al., 1977}) \quad 5.2.3.3$$

where $m_r + n_r$, $r=0, \dots, K-1$ is the set of animals exposed to the risk of failing just before t_r . Both weighted M-H tests place greater emphasis on earlier parts of the survival curves. Clearly, Tarone and Ware test is a less extreme means of weighting the earlier part of the survival curve than Gehan test.

The methods of weighting survival curves suggested in Equations 5.2.3.2 and 5.2.3.3, could be applied to tumor growth curves, in order to test the null hypothesis of equal distribution of tumor volumes in control and treatment arms. One option is to let $m_r + n_r$ be the number of slopes at time t_r , $r=0, \dots, K-1$. $m_r + n_r$ and $\sqrt{m_r + n_r}$ could be used as weights to be introduced in the Formula 5.2.3.1, according respectively to Gehan and Tarone and Ware test, respectively. Again, the former weights tend to place greater importance on the early differences among tumor volume distribution; the latter weights do the same but in a weaker way. The second option is counting the number of days between two consecutive time points t_r and t_{r+1} , $r=0, \dots, K-1$, in which tumor volumes are measured, instead of counting the number of slopes at time t_r , $r=0, \dots, K-1$. The resulting weighted non-parametric two-sample test could be useful in case of unequal distance between time points t_r and

t_{r+1} , $r=0, \dots, K-1$. These two options could be applied jointly, square rooting the product of the size of the sample at time t_r with the number of days between the two consecutive time points t_r and t_{r+1} , for $r=0, \dots, K-1$, according to the following formula:

$$\text{weight}_r = \sqrt{[(m_r + n_r) \cdot (t_{r+1} - t_r)]} \quad 5.2.3.4$$

5.2.4 Stratified non-parametric Two-Sample Tests

A straightforward extension of the statistical procedure given in Section 5.2.1 could be obtained in the presence of strata. Strata could be defined by confounding factors such as animal weight and sex, and by blocking factors such as cage placement within rooms in the animal house and different time periods in which the experiment is performed. Let m ($m=1, 2, \dots, M$) indicate the strata; within each stratum the Mann-Whitney ${}_mW_r$ statistic, its expected value (${}_mE_r$) and variance [$\text{var}_m(W_r)$], $r=0, \dots, K-1$, are computed. This compares like with like. These comparisons within each stratum are then combined to achieve an overall comparison of tumor growth curves by means of a stratified nonparametric two-sample test. The overall test statistic suitable for testing the null hypothesis that tumor volumes are equally distributed in control and treatment arms, for each time t_r , $r=0, \dots, K$, is:

$$\{ \sum_m \sum_r ({}_mW_r - {}_mE_r) \}^2 / \sum_m \text{var}(\sum_r {}_mW_r) \quad 5.2.4.1$$

for $r=0, \dots, K-1$ and $m=1, 2, \dots, M$. This statistic is asymptotically distributed as chi-squared with one degree of freedom.

Because animals compared in each stratum are matched for some characteristics, the stratified analysis tends to increase the power of the test by increasing the difference at the numerator and by decreasing the variance (denominator). In *in vivo* antitumor activity studies, the stratified analysis should be preferred in the following instances:

- when making comparison between control and treatment arms, we need to ensure as much as possible that the differences observed in tumor volumes between arms is due to the treatments and not since the two groups of animals have inherently differing prognoses. We thus may wish to adjust for prognostic variables, such as animals' sex or tumor weight at time t_0 . After the strata has been defined by different modalities of a prognostic variable or by combination of different levels of two or more such variables, a stratified analysis should be performed
- in an randomized complete block (RCB) design, typically each observation can be classified by two factors, one "fixed" (usually called the treatment that is deliberately varied and is of scientific interest) and the other "random" (which may be called a "block" or replicate), which is of no scientific interest but which could cause noise if not removed in the statistical

analysis. Randomization is done separately within each block. Common examples of blocks in *in vivo* activity studies are: batches of animals, which could be of a slightly different age or weight with possible differences in the batches of diet; cages, which could be placed at different position in the enclosure and the period of time over which the experiment is executed. If a RCB design is chosen, a stratified analysis is mandatory. RCB designs are more powerful, have higher external validity, are less subject to bias, and produce more reproducible results than the completely randomized designs typically used in research (Festing, 2014)

- in multicenter preclinical studies, multiple independent laboratories collaboratively conduct a research experiment using a shared protocol. The use of a multicenter design in *in vivo* experimentation is a recent and innovative approach. This kind of design has been suggested as a method to improve reproducibility, generalizability and clinical translation of preclinical work (Boltze *et al.*, 2016; Chamuleau *et al.*, 2018; Dirnagl *et al.*, 2012; Langley *et al.*, 2017; Maertens *et al.*, 2017; O'Brien *et al.*, 2013). Each participating center is a 'block' and a stratified analysis is required
- a systematic review and meta-analysis of all available evidence from preclinical studies should be performed before clinical trials begin. Beneficial evidence obtained from a single laboratory or a single model or a species is probably not sufficient (van der Worp *et al.*, 2010). If raw data is available, each *in vivo* antitumor activity study should be treated as a 'block' and a stratified analysis is required.

It may so happen that treatment effect varies widely among strata. This may be evidence against pooling across strata to compute an overall test of difference in tumor volume distribution. Testing homogeneity across strata will be treated in Section 5.5.

5.2.5 Paired data

In some experiments mice may be grafted bilaterally, obtaining tumor growth on both flanks of the rodent. In this case, one has clustered data (i.e. data obtained from correlated observations in a single rodent). In the presence of clustering, true p-values will be overestimated or underestimated depending on the relationship between paired data and the confidence interval will be too narrow when using standard statistical procedures. In the case of balanced designs (i.e. the same number of replicates per rodent within each arm), a possible approach could be to rank all right tumor volume values and all left tumor volume values, separately, calculate the average rank or rank sum

within a cluster (i.e the rodent), and use our family of statistical tests on the cluster-specific mean rank or rank sum, using the cluster as the unit of analysis.

This approach is inappropriate for unbalanced designs where paired data should be weighted more heavily than single data. In this case, a possible approach is to modify the variance of the Mann-Whitney statistic or the related Wilcoxon rank-sum statistic. Rosner *et al.* (1999) proposed a Mann-Whitney statistics for clustered data which corrects the variance of the test statistic for four types of intracluster correlation coefficients. These correlation coefficients could be estimated by experimental data. Rosner *et al.* (1999) did not provide a large sample theory, instead Rosner *et al.* (2003) proposed a modified Wilcoxon rank-sum statistic under the assumptions that all observations from the same cluster belong to the same treatment arm, observations within any cluster are exchangeable, and that the intracluster dependence does not vary across groups. They derived the asymptotic mean and variance that accommodates unequal cluster sizes and possible stratification.

But above all, resampling techniques, such as those applied in SAS MACRO programs reported in Appendix E, could be used to obtain exact p-values and estimate credible intervals in presence of paired data (refer to Section 5.3.3).

5.3 Statistical power and sample size determination

5.3.1 Location shift model

In *in vivo* experiments one of the most important issues is to determine the number of animals that need to be accrued in order to detect a difference, relevant at least from a biological viewpoint, between control and active arms, with high probability. This issue is fundamental for ethical, scientific and economic reasons. Unfortunately, a totally satisfactory solution has not yet been found (Festing, 2018). As R.A Fisher taught us (Fisher, 1935), the alternative hypothesis (i.e. tumor volumes are differently distributed in control and treatment arms in at least one time point t_r , $r=1,...,K$) is inexact. What this means is that because there are an infinite number of ways in which tumor volumes can be differently distributed, the distribution of test statistic under the alternative hypothesis is not exactly defined. It is necessary to reduce the myriad of possibilities within the alternative space to know precisely the distribution of the test statistic. It is worse than this. The task of determining the exact distribution of a test statistic under an alternative hypothesis is usually more difficult than under the null hypothesis. Generally, the null hypothesis simplifies the problem both mathematically and conceptually.

For instance, in survival analysis, the hypothesis that survival distributions of control and active arms are the same implies that deaths are distributed randomly across control and active

arms, which gives the distribution of deaths occurring at a specific time point a relatively simple distribution (i.e. hypergeometric distribution). Based on this distribution, the distribution of the log-rank statistic can be determined easily. Furthermore, if we reduce the myriad of possibilities within the alternative space to the proportional hazard assumption, it can be shown that the log-rank statistic is distributed asymptotically normal with a mean equal to the log hazard ratio and a variance equal to $[d \cdot p_1 \cdot (1-p_1)]^{(-1)}$, where d is the number of deaths and p_1 is the proportion of experimental units randomly assigned to the control arm (Schoenfeld, 1981; Schoenfeld, 1983).

In our antitumor activity studies, one way to reduce the myriad of possibilities within the alternative hypothesis space, and to help gain a quantitative evaluation about the distribution of the test statistic, is to make the simplifying assumption of an additive effect. The model of an additive treatment effect (i.e. location shift model) assumes that animals do not interfere with each other and the administration of the experimental treatment decreases the slopes $g_{r,i}$ computed from the control arm, by a constant amount $\Delta_r > 0$, for all animals in the time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$. The parameter Δ_r then measures the effect of the experimental treatment in the time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$. The vector of parameters $\underline{\Delta} = (\Delta_0, \dots, \Delta_{K-1})$ to estimate is the additive treatment effect. Negative values of Δ_r , $r=0, \dots, K-1$, correspond to the possibility that the experimental treatment has a detrimental effect on tumor growth curves. The case $\Delta_r = 0$, for all $r=0, \dots, K-1$ corresponds to the case of no treatment effect. The case $\Delta_r = \Delta$, for all $r=0, \dots, K-1$, corresponds to the case of constant treatment effect in each time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$.

Let $X_{r,1}, \dots, X_{r,m(r)}$ be the sample of slopes computed from the experimental arm, identically and independently distributed with probability distribution function $T_r(x)$, for all $x \in (-\infty, +\infty)$, at time t_r , $r=0, \dots, K-1$. Let $Y_{r,1}, \dots, Y_{r,n(r)}$ be the sample of slopes computed from the control arm, identically and independently distributed with probability distribution function $C_r(x)$, for all $x \in (-\infty, +\infty)$, at time t_r , $r=0, \dots, K-1$. Under the assumption of an additive effect, $T_r(x) = C_r(x) - \Delta_r$, for all $x \in (-\infty, +\infty)$, $\Delta_r \in (-\infty, +\infty)$ and $r=0, \dots, K-1$. Let $T = \sum_r T_r$, $C = \sum_r C_r$, $r=0, \dots, K-1$ and $\underline{\Delta}$ the vector $(\Delta_0, \dots, \Delta_{K-1})$. If $\Delta_r > 0$ for all $r=0, \dots, K-1$, C_r and C are random variables stochastically larger than T_r and T , respectively.

The power function of the one-tailed test defined in Section 5.2.1 is given by

$$\Pi(\underline{\Delta}) = P_{\underline{\Delta}}(\sum_r W_r \geq c) \quad 5.3.1.1$$

where W_r is the Mann-Whitney statistic computed in the time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$, c is the critical value determined so that the area to the right of c is equal to the significance level of the one-sided test and finally, $P_{\underline{\Delta}}$ indicates that the probability is calculated for the location shift model. Since $\underline{\Delta} = \underline{0}$ corresponds to the case of no treatment effect, the value $\Pi(\underline{0})$ is just the significance

level of the one-sided test. In case of two-tailed tests (i.e. a detrimental effect of the experimental treatment on tumor growth curves is assessed), the power function $\Pi(\underline{\Delta})$ is defined over the range $(-\infty, +\infty)$ and the significance level is divided in half to account for the two tails of the sampling distribution of the test statistic, under the null hypothesis.

Lemma 5.3.1.1, that is reported and demonstrated in the Lehmann's book (2006), will be used to prove qualitative properties of $\Pi(\underline{\Delta})$ in the following theorems.

Lemma 5.3.1.1: Let $K = 1$. The power function $\Pi(\underline{\Delta})$ defined by Equation 5.3.1.1 is a nondecreasing function of $\underline{\Delta}$.

Proof. If $K = 1$, the vector of parameters $\underline{\Delta}$ estimating the additive treatment effect reduce to the scalar Δ_0 . Let $0 \leq \Delta_0^\alpha < \Delta_0^\beta$. Let $Y_{0.1}, \dots, Y_{0.n(r)}$ be the sample of slopes computed from the control arm, independently and identically distributed with distribution $C_0(x)$ and $X_{0.1}, \dots, X_{0.m(r)}$ be the sample of slopes computed from the experimental arm, independently and identically distributed with distribution $T_0(x) = C_0(x) - \Delta_0^\alpha$. If $S_{0.i} = X_{0.i} - (\Delta_0^\beta - \Delta_0^\alpha)$, $i=1, \dots, m(r)$, the distribution of the random variables $S_{0.1}, \dots, S_{0.m(r)}$ is $C_0(x) - \Delta_0^\beta$. It follows that $\Pi(\Delta_0^\alpha) = P(W_{YX} \geq c)$ and $\Pi(\Delta_0^\beta) = P(W_{YS} \geq c)$, where W_{YX} denotes the number of pairs $(Y_{0.j}, X_{0.i})$ with $Y_{0.j} > X_{0.i}$ and W_{YS} the number of pairs $(Y_{0.j}, S_{0.i})$ with $Y_{0.j} > S_{0.i}$. Because $S_{0.i} < X_{0.i}$ for all $i=1, \dots, m(r)$, it is seen that $W_{YX} \leq W_{YS}$ and hence that $\Pi(\Delta_0^\alpha) \leq \Pi(\Delta_0^\beta)$ as was to be proved.

Lemma 5.3.1.1 says that if there is just one time interval, then the power of the significance test increases monotonically with respect to the size of the additive treatment effect.

Theorem 5.3.1.1: Let $K \geq 1$. Let $\underline{\Delta} = (\Delta_0, \dots, \Delta_{K-1})$, the vector of parameters estimating the additive treatment effect. For each $\Delta_r > 0$, $r=0, \dots, K-1$, the power function $\Pi(\underline{\Delta})$ defined by Equation 5.3.1.1 is a nondecreasing function of Δ_r .

Proof: If $K = 1$, the proof is furnished by Lemma 5.3.1.1. Let $K > 1$ and $\underline{\Delta}_s^\alpha$ and $\underline{\Delta}_s^\beta$ the vectors of parameters, respectively $(\Delta_0^\alpha, \dots, \Delta_{K-1}^\alpha)$ and $(\Delta_0^\beta, \dots, \Delta_{K-1}^\beta)$, estimating the additive treatment effect, with $\Delta_r^\alpha \geq 0$ and $\Delta_r^\beta \geq 0$ for all $r=0, \dots, K-1$, $0 \leq \Delta_s^\alpha < \Delta_s^\beta$ for a specific s , $0 \leq s \leq K-1$, and $\Delta_r^\alpha = \Delta_r^\beta$ for all $r=0, \dots, K-1$, $r \neq s$. Using the same argument of lemma 5.3.1.1, one can prove that $W_r^\alpha \leq W_r^\beta$, for all $r=0, \dots, K-1$, where

a) W_r^α denotes the number of pairs $(Y_j^{\alpha,r}, X_i^{\alpha,r})$ with $Y_j^{\alpha,r} > X_i^{\alpha,r}$, where $Y_j^{\alpha,r}$, $j=1, \dots, n(r)$, is the slope calculated from the control arm, at time r , $r=0, \dots, K-1$, and $X_i^{\alpha,r}$, $i=1, \dots, m(r)$, is the slope calculated from the experimental arm, at time r , $r=0, \dots, K-1$, under the additive treatment effect Δ_r^α

b) W_r^β denotes the number of pairs $(Y_j^{\beta,r}, X_i^{\beta,r})$ with $Y_j^{\beta,r} > X_i^{\beta,r}$, where $Y_j^{\beta,r}$, $j=1, \dots, n(r)$, is the slope calculated from the control arm, at time r , $r=0, \dots, K-1$, and $X_i^{\beta,r}$, $i=1, \dots, m(r)$, is the slope calculated from the experimental arm, at time r , $r=0, \dots, K-1$, under the additive treatment effect Δ_r^β .

Therefore $(\sum_r W_r^\alpha) \leq (\sum_r W_r^\beta)$, $r=0, \dots, K-1$ and hence $\prod(\underline{\Delta}^\alpha) \leq \prod(\underline{\Delta}^\beta)$ as was to be proved.

Theorem 5.3.1.1 says that the power of the significance test increases monotonically with respect to the size of the additive treatment effect in a specified time interval, if the size in all the other time intervals is held constant.

Theorem 5.3.1.2: Let $K \geq 1$. Let $\underline{\Delta} = (\Delta_0, \dots, \Delta_{K-1})$, the vector of parameters estimating the additive treatment effect. The power function $\prod(\underline{\Delta})$, defined by Equation 5.3.1.1, is a nondecreasing function of $\underline{\Delta}$.

Proof. If $K = 1$, the proof is furnished by lemma 5.3.1.1. Let $K > 1$ and $\underline{\Delta}^\alpha$ and $\underline{\Delta}^\beta$ the vectors of parameters, respectively $(\Delta_0^\alpha, \dots, \Delta_{K-1}^\alpha)$ and $(\Delta_0^\beta, \dots, \Delta_{K-1}^\beta)$, estimating the additive treatment effect, with $0 \leq \Delta_r^\alpha \leq \Delta_r^\beta$ for all $r=0, \dots, K-1$ and $0 \leq \Delta_s^\alpha < \Delta_s^\beta$ for at least one s , $0 \leq s \leq K-1$. Using the same argument of lemma 5.3.1.1, one can prove that $W_r^\alpha \leq W_r^\beta$, for all $r=0, \dots, K-1$ (refer to the previous theorem for the meaning of W_r^α and W_r^β notation). Therefore $(\sum_r W_r^\alpha) \leq (\sum_r W_r^\beta)$, $r=0, \dots, K-1$, and hence $\prod(\underline{\Delta}^\alpha) \leq \prod(\underline{\Delta}^\beta)$ as was to be proved.

Theorem 5.3.1.2 says that the power of the significance test increases monotonically if the size of the additive treatment effect in every time interval is either increased or held constant, i.e. the size is not decreased in any time interval.

Corollary 5.3.1.1: if $\prod(\underline{0}) = \alpha$ is the significance level of the family of one-tailed tests defined in Section 5.2.1, it follows from Theorem 5.3.1.2 that for every continuous function $C_r(x)$, for all $x \in (-\infty, +\infty)$, at time t_r , $r=0, \dots, K-1$

$$\prod(\underline{\Delta}) \geq \alpha, \text{ for all } \underline{\Delta} \geq 0 \quad 5.3.1.2$$

Corollary 5.3.1.1 says that the power of the significance test is always greater than or equal to its false-positive rate, provided that the treatment effect is greater than or equal to zero in every time interval. A test whose power against a class of alternatives never falls below the significance level is said to be unbiased against these alternatives. The inequality 5.3.1.2 thus shows that our family of statistical tests is unbiased against the location shift alternatives.

Corollary 5.3.1.2: Let $K \geq 1$. Let $\underline{\Delta} = (\Delta_0, \dots, \Delta_{K-1})$, the vector of parameters estimating the additive treatment effect, and $\Delta_r = \Delta$, for all $r=0, \dots, K-1$ (i.e. constant additive treatment effect in each time

interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$. It follows from Theorem 5.3.1.2 that the power function $\Pi(\underline{\Delta})$, defined by Equation 5.3.1.1, is a nondecreasing function of the constant additive treatment effect $\underline{\Delta}$.

The previous properties describe how the power function depends qualitatively on the treatment effect $\underline{\Delta}$. When determining sample size, it is useful to describe how the power function depends on the number of slopes and the number of time intervals of our antitumor activity studies. For this aim, the following theorem is crucial, although it depends on the following strong assumption:

- slopes in different intervals, in the same animal, are distributed independently (i.e. $\rho_{i,j} = 0$, where $\rho_{i,j}$ is the correlation between slopes at time intervals $[t_i, t_{i+1}]$ and $[t_j, t_{j+1}]$, $i \neq j$, $i, j = 1$ to $0, \dots, K-1$).

This assumption (i.e. independence between time-intervals) will often not be valid in practice.

Theorem 5.3.1.3: Let $K \geq 2$. Let $\underline{\Delta} = (\Delta_0, \dots, \Delta_{K-2})$ and $\underline{\Delta}^* = (\Delta_0, \dots, \Delta_{K-1})$ the vector of parameters estimating the additive treatment effect in the time intervals $[t_r, t_{r+1}]$, $r=0, \dots, K-1$, with $\Delta_r = \Delta$, for all $r=0, \dots, K-1$, $\Delta > 0$ (i.e. constant treatment effect in each time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$). Let the continuous distribution $C_r(x) = C_0(x)$ and $T_r(x) = C_0(x) - \Delta$, $x \in (-\infty, +\infty)$, for all $r=0, \dots, K-1$. Finally, assuming the independence between time-intervals and that the probability of a slope computed from the experimental arm, which is smaller than a slope computed from the control arm, is less than one (i.e. uncertain event), it follows that:

1. the power function $\Pi(\underline{\Delta})$ defined by equation 5.3.1.1 is an increasing function of the number of slopes computed from the experimental and control arm
2. $\Pi(\underline{\Delta}) < \Pi(\underline{\Delta}^*)$

Proof. Let $Y_{r,0}, \dots, Y_{r,n(r)}$ be the sample of slopes computed from the control arm, independently and identically distributed with distribution $C_0(x)$ and let $X_{r,0}, \dots, X_{r,m(r)}$ be the sample of slopes computed from the experimental arm, independently and identically distributed with distribution $T_0(x) = C_0(x) - \Delta$, $r=0, \dots, K-1$. By definition, the test statistic W_r denotes the number of pairs $(Y_{r,i}, X_{r,j})$ with $Y_{r,i} > X_{r,j}$. From the fact that $C_r(x)$ and $T_r(x)$ do not vary across the time intervals and assuming the independence between time-intervals, the probability $P(Y_{r,i} > X_{r,j})$ is constant across the time intervals with a value in the interval $(0,1)$. Hence the test statistic $\sum_r W_r$ could be seen as the sum of independent and identically distributed Bernoulli trials with probability $P(Y_{r,i} > X_{r,j})$. The number of Bernoulli trials are equal to the number of pairs $(Y_{r,i}, X_{r,j})$, $r=0, \dots, K-1$. Based on the properties of the exact binomial distribution, the theorem is proved.

Corollary 5.3.1.3: Let $K \geq 2$. Let $\underline{\Delta} = (\Delta_0, \dots, \Delta_{K-2})$ and $\underline{\Delta}^* = (\Delta_0, \dots, \Delta_{K-1})$ the vector of parameters estimating the additive treatment effect in the time intervals $[t_r, t_{r+1}]$, $r=0, \dots, K-1$. Let the continuous distribution $C_r(x) = C_0(x)$ and $T_r(x) = C_0(x) - \Delta_r$, $x \in (-\infty, +\infty)$, $r=0, \dots, K-1$, $\Delta_r \geq 0$ for all $r=0, \dots, K-1$, $\Delta_r > 0$ for at least one $r=0, \dots, K-1$. If $\Delta_{K-1} \geq \Delta_r$ for all $r=0, \dots, K-1$ then $\Pi(\underline{\Delta}) < \Pi(\underline{\Delta}^*)$

Proof. Let $\underline{\Delta}^s = (\Delta_0, \dots, \Delta_{K-2})$, $\Delta_j = \Delta_{K-1}$ for all $j=1, \dots, K-2$. From Theorem 5.3.1.3 it deduced that $\Pi(\underline{\Delta}^*) > \Pi(\underline{\Delta}^s)$. From Theorem 5.3.1.2 it deduced that $\Pi(\underline{\Delta}^s) \geq \Pi(\underline{\Delta})$ and hence $\Pi(\underline{\Delta}^*) > \Pi(\underline{\Delta})$ as was to be proved.

Theorem 5.3.1.3 and its generalization, Corollary 5.3.1.3, say that, by assuming the independence between time-intervals, more animals or time intervals will always give us more power. However, it is important to note that the assumption of independence between time-intervals in the same animal may be unrealistic.

5.3.2 Asymptotic power

Under the assumption of an additive treatment effect, the theorems on the power function $\Pi(\underline{\Delta})$ proved in the preceding section were only qualitative. To calculate the power quantitatively, it is necessary to know the distribution of the ranks when the function $C_r(x)$, $x \in (-\infty, +\infty)$, at time t_r , $r=0, \dots, K-1$, is continuous and $\Delta_r \geq 0$, for all $r=0, \dots, K-1$. Useful results can be obtained from the normal approximation to the power. Let $[t_r, t_{r+1}]$ be a specific time interval, $r=0, \dots, K-1$. Under the alternative hypothesis, the distribution of the Mann-Whitney W_r statistic at time t_r tends to a normal distribution as m_r and n_r tend to infinity, for any continuous distributions $C_r(x)$ and $T_r(x) = C_r(x) - \Delta_r$, $x \in (-\infty, +\infty)$ and $\Delta_r > 0$, for which

$$0 < P(X < Y) < 1 \quad 5.3.2.1$$

where X and Y indicate two independent random variables with distribution T_r and C_r , respectively. If the probability of a slope, computed from experimental arm, being smaller than a slope, computed from control arm, is equal to 0 (impossible event) or 1 (certain event), the distribution T_r lies entirely to the right or to left of the distribution C_r , respectively. In either case, W_r statistic reduces to a constant. To prove asymptotic normality of W_r statistic under the assumption of an additive effect, refer to the example 20 reported in the Appendix of Lehmann's book (2006).

Application of the normal approximation requires the expectation and variance of the W_r statistic under the hypothesis of an additive treatment effect. The expectation $E(W_r)$ depends on the probability $p_1 = P(X < Y)$. It can be proved that

$$E(W_r) = m_r \cdot n_r \cdot p_1 \quad 5.3.2.2$$

As a check, consider the case $\Delta_r = 0$. The variables X and Y then have the same distribution, so that $P(X < Y) = P(Y < X)$. From the fact that this common distribution is continuous, $P(X = Y) = 0$, and hence $p_1 = 1/2$. The resulting value $E(W_r) = m_r \cdot n_r / 2$ agrees with that given by Formula 5.2.1.2. The variance of the W_r statistic depends, besides p_1 , on the two following quantities:

$$p_2 = P(X < Y \text{ and } X < Y') \quad 5.3.2.3$$

where X , Y and Y' indicate three independent random variables, X with distribution T_r , and Y and Y' each with distribution C_r

$$p_3 = P(X < Y \text{ and } X' < Y) \quad 5.3.2.4$$

where X , X' and Y indicate three independent random variables, X and X' with distribution T_r , and Y with distribution C_r .

With this notation, it can be proved that

$$\text{var}(W_r) = m_r \cdot n_r \cdot p_1 \cdot (1-p_1) + m_r \cdot n_r \cdot (n_r-1) \cdot (p_2 - p_1^2) + m_r \cdot n_r \cdot (m_r-1) \cdot (p_3 - p_1^2) \quad 5.3.2.5$$

Again, as a check, consider the case $\Delta_r = 0$. Then p_2 becomes the probability that of three independent random variables X , Y and Y' with the same continuous distribution, X is the smallest. Since each of the three is equally likely to be the smallest, then $p_2 = 1/3$ and by the same argument that $p_3 = 1/3$. As previously proved, p_1 is equal to $1/2$. The resulting value $\text{var}(W_r) = m_r \cdot n_r \cdot (m_r + n_r + 1) / 12$ agree with that given by Formula 5.2.1.3. To prove Formula 5.3.2.2 and 5.3.2.5, refer to the example 5 reported in the Appendix of Lehmann's book (2006). It could be useful to note that $p_2 = p_3$, in the case where C_r is symmetric.

In principle, the asymptotic power can be computed for any alternative $C_r(x)$ and $T_r(x) = C_r(x) - \Delta_r$, $x \in (-\infty, +\infty)$ and $\Delta_r > 0$. In other terms, once C_r , T_r and Δ_r have been specified, the values of p_1 , p_2 and p_3 are determined and the expectation $E(W_r)$ and variance $\text{var}(W_r)$ can be computed using Formula 5.3.2.2 and 5.3.2.5. Note that the computations of p_2 and p_3 are typically more involved than that of p_1 . However, there are some important cases for which analytical forms are available:

- Normal distribution [i.e. $C_r(x) = N(\zeta, \sigma^2)$ and $T_r(x) = N(\zeta - \Delta_r, \sigma^2)$, $x \in (-\infty, +\infty)$]:

$p_1 = \Phi[\Delta_r / (\sigma \cdot 2^{(1/2)})]$, where $\Phi(z)$ is the Standard Normal Cumulative Distribution Function

$p_2 = p_3 = P(Z \leq z, Z' \leq z)$, where Z and Z' are both normal with mean zero and unit variance and correlation coefficient equal to 0.5, z is equal to $[\Delta_r / (\sigma \cdot 2^{(1/2)})]$

For instance, if $\Delta_r = 5$ and $\sigma^2 = 32$, so that $[\Delta_r / (\sigma \cdot 2^{(1/2)})] = 0.625$, p_1 , p_2 and p_3 are equal to 0.734, 0.600 and 0.600, respectively

- Standard uniform distribution, defined in the interval $(-1/2, 1/2)$, with $0 \leq \Delta_r \leq 1$:

$$p_1 = (1/2) + \Delta_r \cdot (1 - \Delta_r/2)$$

$$p_2 = p_3 = (1/3) + \Delta_r - (\Delta_r^3)/3$$

- Standard double exponential distribution:

$$p_1 = 1 - \{ (1/2) \cdot [1 + (\Delta_r/2)] \cdot \exp(-\Delta_r) \}$$

$$p_2 = p_3 = 1 - [(7/12) + (\Delta_r/2)] \cdot \exp(-\Delta_r) - [(1/12) \cdot \exp(-2 \cdot \Delta_r)]$$

- Standard exponential distribution:

$$p_1 = 1 - [(1/2) \cdot \exp(-\Delta_r)]$$

$$p_2 = 1 - [(2/3) \cdot \exp(-\Delta_r)]$$

$$p_3 = 1 - \exp(-\Delta_r) + [(1/3) \cdot \exp(-2 \cdot \Delta_r)]$$

Because the distribution of the W_r statistic is discrete, a continuity correction should be applied to compute the asymptotic power. If the hypothesis $\Delta_r = 0$ is rejected when $W_r \geq c$, the asymptotic power with continuity correction is

$$P_{\Delta_r'}(W_r \geq c) = 1 - \Phi\{ [c - (1/2) - E(W_r)] / [(\text{var}(W_r))^{(1/2)}] \} \quad 5.3.2.6$$

where $E(W_r)$ and $\text{var}(W_r)$ are the expectation and variance computed in the specific time interval $[t_r, t_{r+1}]$, under the alternative hypothesis $\Delta_r > 0$.

Previous results refer to a specific time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$. Based on these preliminary results, the asymptotic power of the family of statistical tests, defined in Section 5.2.1, could be computed. First of all, the following theorem is proved:

Theorem 5.3.2.1: Let $K \geq 1$. Let $\underline{\Delta} = (\Delta_0, \dots, \Delta_{K-1})$ be the vector of parameters estimating the additive treatment effect. Let W_r be the Mann-Whitney statistic computed in the time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$, under the hypothesis of an additive treatment effect. Then the test statistic $\sum_r W_r$, $r=0, \dots, K-1$, is asymptotically normally distributed.

Proof. If $K = 1$, the proof is furnished by the previous consideration reported in this section. Let $K > 1$. For each r , $r=0, \dots, K-1$, W_r is asymptotically normally distributed with expectation and variance given by Formula 5.3.2.2 and 5.3.2.5, respectively. Because it is the sum of correlated normal random variables, $\sum_r W_r$, $r=0, \dots, K-1$, is asymptotically normally distributed. Its expectation is the sum of expectation of each W_r , $r=0, \dots, K-1$. Its variance is the sum of the elements of the covariance matrix.

Hence, if the hypothesis $\underline{\Delta} = \underline{0}$ is rejected when $\sum_r W_r \geq c$, the asymptotic power of the test against any fixed alternative hypothesis $T_r(x) = C_r(x) - \Delta_r$, $x \in (-\infty, +\infty)$ and $\Delta_r \geq 0$, $T_r(x) \neq C_r(x)$ for at least one r , $r=0, \dots, K-1$, is

$$P_{\underline{\Delta}}(\sum_r W_r \geq c) = 1 - \Phi\{ [c - ((1/2) + \sum_r E(W_r))] / [(\text{var}(\sum_r W_r))^{(1/2)}] \} \quad 5.3.2.7$$

where $E(W_r)$ is the expectation computed in each time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$, and $\text{var}(\sum_r W_r)$ is the sum of the elements of the covariance matrix, under the alternative hypothesis of an additive treatment effect.

Asymptotic power is calculated in the following example: suppose that an *in vivo* experiment comparing a control and an active arm is performed, and the following assumptions are made:

- 10 rodents are assigned to both arms
- tumor volumes are assessed at consecutive time intervals from randomization. The number of time intervals vary between 3 and 6
- tumor volumes increase over time according to an exponential growth in both arm. The growth rate of the control arm is equal to 0.07
- slopes are distributed normally with standard deviation equal to 0.060 in both arms
- slopes in different intervals, in the same animal, are distributed independently (i.e. $\rho_{i,j} = 0$, where $\rho_{i,j}$ is the correlation between slopes at time intervals $[t_i, t_{i+1}]$ and $[t_j, t_{j+1}]$, $i \neq j$, $i, j = 1$ to $0, \dots, K-1$). Hence, the following equality holds: $\text{var}(\sum_r W_r) = \sum_r \text{var}(W_r)$, $r=0, \dots, K-1$.

In Table 5.3.2.1 asymptotic power is reported for different treatment effects Δ (i.e. difference in growth rate between control and active arm). Treatment effect is the same in different time intervals. The test is two-tailed with type I error equal to 0.05.

Δ (day ⁻¹)	N° of time intervals			
	3	4	5	6
0.01	9.2	10.8	12.4	14.0
0.02	23.1	29.3	35.3	41.1
0.03	44.6	56.2	66.0	74.0
0.04	68.3	80.9	88.9	93.8
0.05	86.7	94.8	98.1	99.3

Table 5.3.2.1 Asymptotic power for different treatment effects. Power is given as percent

The values in the Table 5.3.2.1 show monotonic positive relationships between treatment effects Δ and power and between number of time intervals and power. This result is a direct illustration of Theorem 5.3.1.3 and its generalization, Corollary 5.3.1.3, as the assumption of independence between time intervals was used. Caution should be taken in generalizing this result because the

assumption of independence between time intervals in the same animal will often not be met in real cases.

A SAS MACRO program to compute asymptotic power and sample size of an *in vivo* experiment is reported in Appendix E. Assumptions made by the SAS program are the following:

- slopes are distributed normally
- slopes in different intervals in the same animal are distributed independently. Hence, the following equality holds: $\text{var}(\sum_r W_r) = \sum_r \text{var}(W_r)$, $r=0,\dots,K-1$. This strong assumption does not usually hold for real cases
- distribution of slopes and treatment effect does not vary over time intervals

5.3.3 Other approximations and exact distribution of test statistic

The distribution of the test statistic, under the null and alternative hypothesis, has been previously studied asymptotically. However, for ethical, economic and scientific reasons, *in vivo* experiments use small samples. To better approximate the distribution of the Mann-Whitney statistic and, hence, to better approximate the distribution of the test statistic proposed for studying tumor growth curves, an Edgeworth series could be used (Hall, 1992). Formulas are provided by Fix *et al.* (1955). They also provide a comparison of the accuracy of normal approximation with that derived from Edgeworth series. The approximations are studied further by Verdooren (1963) and Bickel (1974). The corresponding problem in the presence of ties is investigated by Klotz (1966).

The exact distribution of the Mann-Whitney statistic, sometimes substituted by that of the equivalent Wilcoxon rank-sum statistic, is usually tabulated in books about non-parametric statistical methods. For example, Lehmann's book (2006) tabulates the cumulative distribution of the Wilcoxon rank-sum statistic. The computation of the exact distribution of rank statistic under alternative hypotheses is difficult, and only such few computations have been carried out (Bell *et al.*, 1966; Haynam *et al.*, 1966; Milton, 1970). More than this, the computation of the exact distribution of the proposed test statistic, sum of dependent exact distributions of Mann-Whitney statistic, have never been carried out. Bootstrap methods and resampling techniques could be easily used to compute the exact distribution. In Appendix E, two SAS MACRO programs compute both the asymptotic and exact p-values. The asymptotic p-value is calculated under the strong assumption that slopes in different intervals, in the same animal, are distributed independently.

5.4 An example of statistical analysis of tumor growth curves

The new family of statistical tests was applied to different *in vivo* experiments performed at the Oncology Department, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan (Italy). As an

example, the antitumor activity of the combination of trabectedin (Yondelis®) with the PPAR γ agonist pioglitazone was evaluated in various patient-derived myxoid liposarcoma xenografts (PDXs), characterized by different sensitivity to trabectedin (Frapolli *et al.*, 2019). The statistical analysis of a specific tumor model, the ML017/ET model, is reported. ML017/ET was obtained from the trabectedin sensitive ML017 PDX, through the exposition at repeated *in vivo* cycles of trabectedin until the acquisition of a resistant phenotype. When tumor burden reached about 300-400 mg, athymic nude mice bearing ML017/ET xenografts were randomized in the following treatment groups:

- control (i.e. placebo arm)
- trabectedin 0.15 mg/kg i.v., every 7 days for three times (q7dx3)
- pioglitazone 150 mg/kg p.o. daily for 28 days
- combination of trabectedin and pioglitazone

Tumor growth was measured using Vernier caliper, and the tumor burden was calculated by the formula: length x (width)² / 2. Mice were sacrificed when tumor burden reached about 1500-2000 mm³. The total number of mice evaluated in different days is reported in Table 5.4.1. Figure 5.4.1 reports the tumor growth curves obtained. Mean tumor volumes and standard errors are shown in the same figure. The table of raw data is reported in Appendix F.

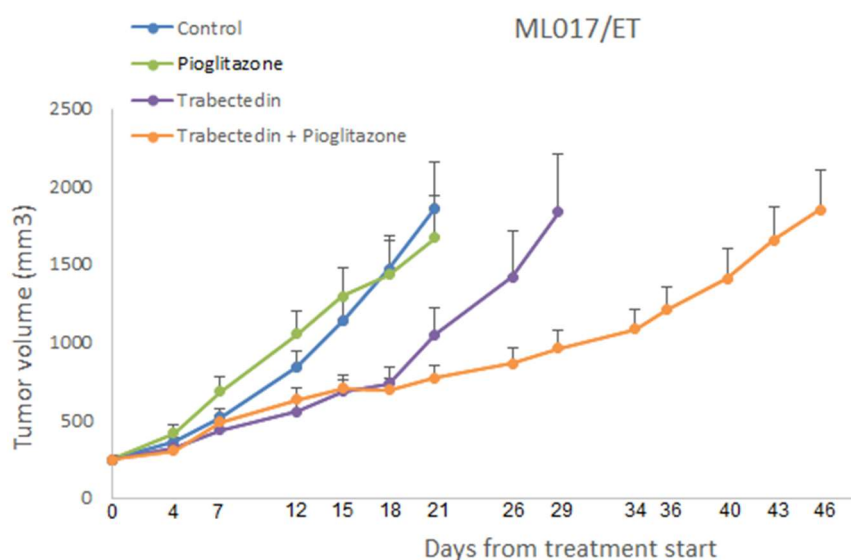


Figure 5.4.1 Tumor growth curves

Arm	Days after treatment													
	0	4	7	12	15	18	21	26	29	34	36	40	43	46
Control	12	12	12	12	12	12	12	0	0	0	0	0	0	0
Pioglitazone	11	11	11	11	11	11	11	0	0	0	0	0	0	0
Trabectedin	11	11	11	11	11	11	11	11	11	0	0	0	0	0
Trab + Pio	11	11	11	11	11	11	11	11	11	11	11	11	11	11

Table 5.4.1 Number of mice at risk

In Table 5.4.2 the two-tailed asymptotic and exact p-values, used to detect differences between arms, are reported. Both types of p-value were calculated using SAS MACRO programs reported in Appendix E. 500 simulations have been used to calculate exact p-values. The level of agreement between the asymptotic and exact tests seems good.

Comparison	χ^2 (Asymptotic p-value)[°]	Exact p-value
Pio vs Ctr	2.051 (0.152)	0.184
Trab vs Ctr	10.669 (0.001)	0.010
Trab vs Pio	3.130 (0.077)	0.080
Trab + Pio Vs Ctr	25.001 (<0.001)	<0.001
Trab + Pio Vs Pio	16.823 (<0.001)	0.002
Trab + Pio. Vs Trab	10.868 (<0.001)	<0.001

Table 5.4.2 Two-tailed asymptotic and exact p-values

Legend 5.4.2: ° Slopes in different intervals in the same animal are assumed to be distributed independently

5.5 Estimating the treatment effect

5.5.1 Introduction

Testing the hypothesis of no treatment effect is a necessary condition in the comparison of the experimental with the control arm. However, after the treatment effect has been detected, it could be relevant to give an estimate of this effect. Assuming the location shift model, the parameter Δ , by which the treatment shifts the tumor growth distribution, is the natural measure of the treatment effect. Let $X_{r,0}, \dots, X_{r,m(r)}$ be the sample of slopes computed from the experimental arm, identically and independently distributed with probability distribution function $T_r(x)$, for all $x \in (-\infty, +\infty)$, at time t_r , $r=0, \dots, K-1$. Let $Y_{r,0}, \dots, Y_{r,n(r)}$ be the sample of slopes computed from the control arm, identically and independently distributed with probability distribution function $C_r(x)$, for all $x \in (-\infty, +\infty)$, at time t_r , $r=0, \dots, K-1$. Under the assumption of an additive effect, $T_r(x) = C_r(x) - \Delta_r$, for all $x \in (-\infty, +\infty)$, $\Delta_r \in (-\infty, +\infty)$, $r=0, \dots, K-1$. Δ_r^{med} and Δ_r^{mean} , respectively the median and mean of the differences between control and experimental slopes, will be proposed as estimators of the parameter Δ_r . In case of constant treatment effect (i.e. $\Delta_r = \Delta$, for all $r=0, \dots, K-1$), estimators Δ_r^{med} and Δ_r^{mean} could be combined for all $r=0, \dots, K-1$ to estimate Δ . The choice of the estimator should always be made before the *in vivo* experiment is performed.

5.5.2 Definition and properties of the estimator Δ_r^{med}

Denote the ordered set of $m_r \cdot n_r$ differences $Y_{r,j} - X_{r,i}$, where $j=1, \dots, n_r$ and $i=1, \dots, m_r$, by $D_{(1)} < D_{(2)} < \dots < D_{m(r) \cdot n(r)}$. The definition of Δ_r^{med} depends on the parity of $m_r \cdot n_r$:

- the product $m_r \cdot n_r$ is even, say $m_r \cdot n_r = 2 \cdot k$. In this case the estimator Δ_r^{med} is the midpoint of the interval $[D_{(k)} - D_{(k+1)}]$. In other words, the estimator is the median of the $D_{(s)}$, $s=1, \dots, 2 \cdot k$

$$\Delta_r^{\text{med}} = (1 / 2) \cdot (D_{(k)} + D_{(k+1)}) \quad 5.5.2.1$$

- the product $m_r \cdot n_r$ is odd, say $m_r \cdot n_r = 2 \cdot k + 1$. Also in this case the estimator Δ_r^{med} is the median of the of the $D_{(s)}$, $s=1, \dots, 2 \cdot k$. The estimator is given by

$$\Delta_r^{\text{med}} = D_{(k+1)} \quad 5.5.2.2$$

Using statistical software such as SAS, Stata or R, it is very easy to calculate the point estimate of Δ_r^{med} in each time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$. A graphical shortcut method of calculating manually the point estimate of Δ_r^{med} is shown by Lehmann's book (2006).

Under the location shift model, the estimator Δ_r^{med} has the following important properties:

- ✓ The distribution of the difference $\Delta_r^{\text{med}} - \Delta_r$ (i.e. the error of the estimator) is independent of Δ_r . Hence, the expected value and the standard error of the Δ_r^{med} estimator are independent from the real value of Δ_r .
- ✓ The estimator Δ_r^{med} of the parameter Δ_r is distributed symmetrically about Δ_r if either of the following two conditions hold:
 1. The distribution C_r is symmetric about some point μ
 2. The two sample sizes are equal, that is, $m_r = n_r$

Under the stated conditions, the estimator Δ_r^{med} of the parameter Δ_r is unbiased (i.e. the expected value of the estimator is exactly Δ_r). Proof of the previous properties are furnished by Lehmann's book (2006). No closed formula is available for the variance of this estimator. Its variance depends on the C_r distribution. Instead, thanks to the following theorem, a credible interval could be associated to the point estimate of Δ_r^{med} :

Theorem 5.5.2.1: For each $k=1, \dots, m_r \cdot n_r$,

$$P_{\Delta=\delta} (D_{(k)} \leq \Delta_r < D_{(k+1)}) = P_{\Delta=0} (W_r = k) \quad 5.5.2.3$$

where $\delta=\Delta_r$. This theorem is fundamental because it links the probability that Δ_r lies between $D_{(k)}$ and $D_{(k+1)}$ to the distribution of the Mann-Whitney statistic under the null hypothesis (i.e. $\Delta=0$). Hence, it permits to define a credible interval around the point estimate of Δ_r using merely the distribution of the Mann-Whitney statistic under the null hypothesis. Proof of this theorem is furnished by Lehmann's book (2006). In appendix E, a SAS MACRO program permits to calculate the credible interval of each Δ_r , $r=0, \dots, K-1$. The SAS program was used to calculate point estimates and credible intervals of tumor growth rates of the example reported in Section 5.4. Tumor growth rates by time interval are reported in Table 5.5.2.1.

5.5.3 Definition and properties of the estimator Δ_r^{mean}

Consider the set of $m_r \cdot n_r$ differences $Y_{r,j} - X_{r,i}$, where $j=1, \dots, n_r$ and $i=1, \dots, m_r$. Then Δ_r^{mean} is defined as the mean value of these differences, or equivalently, as the difference of the mean values of $Y_{r,j}$, $j=1, \dots, n_r$ and $X_{r,i}$, $i=1, \dots, m_r$. The asymptotic properties of this estimator are determined by the central limit theorem. Δ_r^{mean} is an unbiased estimator and its variance could be calculated using the standard deviation of collected slopes {i.e. $[\sigma^2(Y_{r,j}) + \sigma^2(X_{r,i})]^{(1/2)}$ }. Point estimates and 95% confidence intervals of tumor growth rates of Section 5.4 example are reported in Table 5.5.3.1.

Comparison	Tumor growth rate (day ⁻¹)	Time interval (days)							
		0-4	4-7	7-12	12-15	15-18	18-21	21-26	26-29
Pio vs Ctr	Point estimate	3.5	4.9	-1.2	-2.9	-5.5	-3.6	not defined	not defined
	95% credible interval	(-0.3) - 8.0	(-1.2) - 10.1	(-4.4) - 2.9	(-6.7) - 1.8	(-8.1) - (-1.9)	(-6.1) - 0.3		
Trab vs Ctr	Point estimate	-3.1	-1.3	-6.2	-2.0	-5.9	1.8	not defined	not defined
	95% credible interval	(-7.4) - 1.8	(-6.2) - 4.6	(-8.6) - (-3.1)	(-6.4) - 1.6	(-10.4) - (-2.2)	(-1.4) - 7.6		
Trab vs Pio	Point estimate	-6.8	-6.1	-3.9	0.9	-0.9	5.8	not defined	not defined
	95% credible interval	(-10.4) - (-2.5)	(-10.7) - (-0.4)	(-8.2) - (-0.2)	(-3.1) - 4.0	(-8.4) - 2.9	2.5 - 9.5		
Trab + Pio Vs Ctr	Point estimate	-4.8	6.2	-4.9	-6.6	-8.7	-4.8	not defined	not defined
	95% credible interval	(-8.8) - (-0.3)	(-0.1) - 12.3	(-7.0) - (-2.8)	(-10.6) - (-3.0)	(-12.2) - (-4.9)	(-7.6) - (-1.9)		
Trab + Pio Vs Pio	Point estimate	-8.6	1.1	-2.9	-4.2	-3.2	-1.3	not defined	not defined
	95% credible interval	(-12.6) - (-4.3)	(-4.9) - 7.4	(-6.8) - (-0.2)	(-7.1) - (-1.0)	(-6.3) - 0.3	(-3.6) - 1.4		
Trab + Pio Vs Trab	Point estimate	-1.8	7.4	0.9	-4.7	-2.5	-7.0	-2.6	-5.3
	95% credible interval	(-6.0) - 1.9	0.4 - 13.8	(-2.6) - 3.6	(-8.2) - (-1.8)	(-6.1) - 3.7	(-10.8) - (-3.9)	(-6.2) - (-0.4)	(-8.4) - (-2.0)

Table 5.5.2.1 Treatment effects using Δ_r^{med} as estimator

Legend 5.5.2.1: Tumor growth rates are multiplied by 10²

Comparison	Tumor growth rate* (day ⁻¹)	Time interval (days)							
		0-4	4-7	7-12	12-15	15-18	18-21	21-26	26-29
Pio vs Ctr	Point estimate	4.0	4.5	-1.0	-2.0	-4.1	-2.9	not defined	not defined
	95% confidence interval	2.9 - 5.1	3.1 - 5.9	(-1.9) - (-0.2)	(-0.9) - (-3.2)	(-5.3) - (-2.9)	(-3.9) - (-2.0)		
Trab vs Ctr	Point estimate	-2.3	-1.3	-5.4	-1.9	-6.2	3.2	not defined	not defined
	95% confidence interval	(-3.5) - (-1.1)	(-2.7) - (<0.1)	(-6.1) - (-4.7)	(-3.0) - (-0.7)	(-7.7) - (-4.6)	2.2 - 4.3		
Trab vs Pio	Point estimate	-6.3	-5.8	-4.3	0.2	-2.1	6.2	not defined	not defined
	95% confidence interval	(-7.3) - (-5.3)	(-7.1) - (-4.5)	(-5.4) - (-3.3)	(-0.8) - 1.1	(-3.4) - (-0.8)	5.2 - 7.1		
Trab + Pio Vs Ctr	Point estimate	-3.9	5.6	-4.8	-6.2	-7.2	-4.0	not defined	not defined
	95% confidence interval	(-5.0) - (-2.7)	4.0 - 7.3	(-5.3) - (-4.2)	(-7.3) - (-5.1)	(-8.4) - (-6.0)	(-4.9) - (-3.2)		
Trab + Pio Vs Pio	Point estimate	-7.9	1.1	-3.7	-4.2	-3.1	-1.1	not defined	not defined
	95% confidence interval	(-8.9) - (-6.9)	(-0.5) - 2.8	(-4.6) - (-2.8)	(-5.0) - (-3.3)	(-4.0) - (-2.2)	(-1.8) - (-0.4)		
Trab + Pio Vs Trab	Point estimate	-1.6	7.0	0.6	-4.3	-1.0	-7.3	-3.2	-4.3
	95% confidence interval	(-2.7) - (-0.5)	5.4 - 8.5	(-0.1) - 1.4	(-5.1) - (-3.5)	(-2.3) - 0.3	(-8.1) - (-6.4)	(-3.9) - (-2.4)	(-5.3) - (-3.3)

Table 5.5.3.1 Treatment effects using Δ_r^{mean} as estimator

Legend 5.5.3.1: Tumor growth rates are multiplied by 10²

5.5.4 Summary estimators and heterogeneity between time intervals

Δ_r^{med} and Δ_r^{mean} estimators are defined in each time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$. It could be useful to summarize estimates obtained in different time intervals, in a single average estimate. If the assumption that treatment effects in different time intervals are equal is reliable, this average estimate should represent the 'true' treatment effect. For the Δ_r^{mean} estimator, fixed and random effects models for repeated measures could be used to estimate the average treatment effect and the heterogeneity between treatment effects in different time intervals. For the Δ_r^{med} estimator, it is not possible to apply the previous approach because expectation and variance of the estimator's sampling distribution are unknown. Instead, the following estimator could be used:

$$(\sum_r \Delta_r^{\text{med}}) / (K-1) \quad 5.5.4.1$$

Repeated simulations of $\sum_r \Delta_r^{\circ} / (K-1)$, $r=0, \dots, K-1$, where the distribution of the random variable Δ_r° is given by Formula 5.5.2.3, could be used to define credible intervals around the point estimate of Formula 5.5.4.1.

The compatibility between the probability distribution of the estimator defined in 5.5.4.1 and estimates of each Δ_r^{med} , $r=0, \dots, K-1$, obtained in the *in vivo* experiment, could be evaluated using re-sampling techniques and summarized by p-values or graphical display.

5.5.5 Estimating relative effects

In the location shift model the administration of the experimental treatment decreases the slope $g_{r,i}$ of a constant amount Δ_r , for all animals in the time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$. The parameter Δ_r then measures the effect of the experimental treatment in the time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$. Optionally, researchers could prefer to express the effect size as a relative change. For instance, the administration of the experimental treatment reduce the slope $g_{r,j}$ of a relative amount Γ_r , for all animals in the time interval $[t_r, t_{r+1}]$, $r=0, \dots, K-1$. This is no longer an additive model regulated by the parameter Δ_r , this is a multiplicative model regulated by the parameter Γ_r . Theoretical and calculus results of the location shift model could be simply transferred to the multiplicative model using logarithmic scale.

Chapter 6

Discussion

At the beginning of this project, the proposed survey design was very different to the current one. The initial survey design is shown in Figure 6.1. and can be summarized as following steps:

Step 1. Database Search for published clinical studies

A systematic search of Medline and EMBASE databases was to be carried out to identify antitumor activity studies in clinical research (i.e. phase II clinical trials)

Step 2. Identifying CCs

In each of the phase II clinical trials those CCs used as a single agent were to be identified. A random sampling procedure was to be used to retrieve and select CCs

Step 3. Stratified case-control study

CCs were to be classified as cases or controls based on positive (i.e. cases) or negative (i.e. controls) statistical demonstration of antitumor activity. Each case was to be matched to a maximum of three controls using the following variables: a) drug's classification [i.e chemotherapy, targeted therapy] b) single-center or multicenter c) primary endpoint d) single-arm or controlled e) use of randomization. The strata were to be identified by the following variables: a) drug's classification b) primary endpoint

Step 4. Database Search for published animal studies

A systematic search of Medline and EMBASE databases was to be carried out to identify preclinical *in vivo* antitumor activity studies in which CCs were used as monotherapy

Step 5. Evaluation of the quality of statistical design and analysis

For each preclinical *in vivo* experiment the quality of statistical design and analysis was to be assessed using the ad hoc checklist reported in Appendix B. Each checklist item was to be statistically correlated to cases and controls.

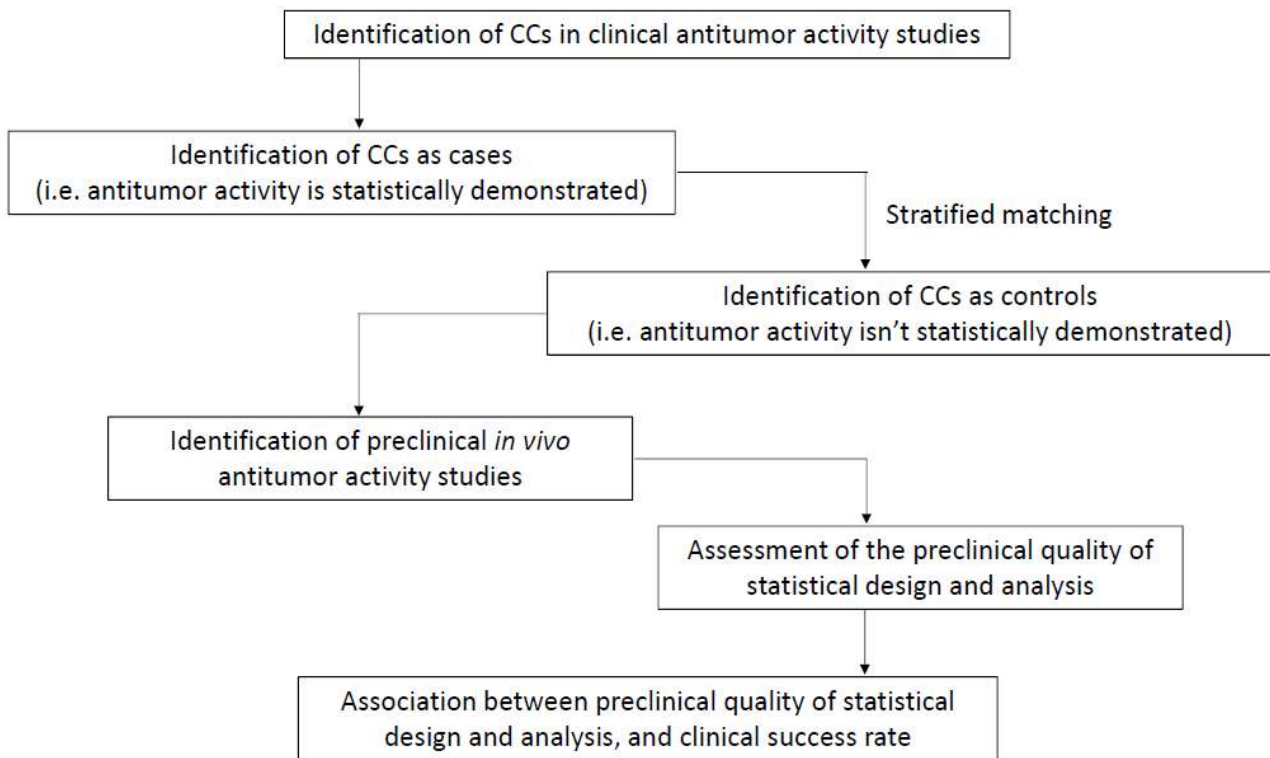


Figure 6.1 Initial survey design

The project's task, using this survey design, was terminated early because the classification of CCs as cases and controls was not reliable. Phase II clinical trials in Oncology are generally single-arm trials. The success rate for single-arm phase II clinical trials in Oncology is strongly influenced by patient selection bias, misclassification error in tumor response and choice of null and alternative hypotheses.

Therefore it was decided to classify CCs as cases (i.e. authorized CCs) and controls (i.e. non-authorized CCs) using the public assessment reports published by the European Medicines Agency and the Food and Drug Administration. But preclinical *in vivo* experiments were not clearly identified in the public assessment reports of the former agency and the public assessment reports of non-authorized CCs were not available from the latter agency. Hence, the project's task, using this modified survey design, was terminated. The survey design defined in Chapter 3 was finally considered.

The first relevant result of this project was that poor methodology and reporting were applied to preclinical *in vivo* antitumor activity experiments. About 3 out of 4 assessed preclinical *in vivo* experiments were performed in the last ten years. Although the survey was limited to EOC models, it is reasonable to assume that the situation was identical for other tumors. These unsatisfactory

results are consistent with those obtained by other surveys not specifically focused in Oncology (Kilkenny *et al.*, 2009).

More than 3 out of 4 preclinical *in vivo* experiments were performed in a single laboratory, using only one species and model. This instills doubts about scientific reproducibility. Poor use of blinding techniques and absence of any reporting of allocation concealment call into question internal validity. Biological and experimental variability was scarcely controlled: if it was applied, randomization was limited to experimental units (i.e rodents); blocks were not used in treatment assignment or in the analysis of longitudinal data. Another interesting finding was that the number of animals allocated to each treatment arm was moderately heterogeneous (IQR: 6-10). This heterogeneity was not supported by a satisfactory justification of sample size. It is worse still. The statistical analysis, particularly that of tumor growth curves, was in the best cases inadequate (e.g. statistical assumptions were not justified, tumor growth curves were analyzed in a single and unjustified time point and confidence intervals were reported in less than half of the preclinical *in vivo* experiments) or even wrong (e.g. use of one-way ANOVA in multiple time points without controlling the statistical error or unequal number of animals assigned to different arms). Poor reporting of the number of animals at risk and lack of reasons of drop-outs render it impossible to fairly estimate biological signal and variability. For ethical reasons, scientific information should be maximized by each *in vivo* experiment. And yet, although factorial designs were greatly used, effects interaction between arms was never formally evaluated and confidence intervals were reported for less than half of *in vivo* experiments. Finally, due to rodent outcome measurement (e.g. tumor volume) and definition (e.g. tumor response, progression and cure) not being standardized, it is very unreliable to make comparison and perform meta-analyses of preclinical *in vivo* antitumor activity experiments.

“The use of animals in biomedical research generates strong emotions, but everyone will surely agree that if they are used the experiments should be properly designed” said Michael Festing (2010) and our integrity requires this. My hope is that the result of this survey could be used by preclinical researchers to improve the quality of statistical design, analysis and reporting of the preclinical *in vivo* antitumor activity experiments.

Particularly, this survey showed that the issue of estimating biological signals was poorly addressed and debated. If a biological signal (e.g. shrinkage of tumor volume at different time points) is formally detected, then its magnitude and pattern should be estimated and questioned. Poor methodological quality implies unreliable estimates of the biological signal. Broadly, it seemed

that researchers were interested to detect the biological signal but were less keen on or unable to debate its magnitude and pattern. In other words, reaching a statistically significant p-value was enough. It is time to overcome this simplification and get to the consequent and crucial step: to critically debate the magnitude and pattern of the biological signal. Better estimating the magnitude of the biological signal is necessary at least for the following reasons:

- to improve preclinical reproducibility and compare *in vivo* antitumor activity experiments
- to perform reliable systematic reviews and meta-analysis. For instance, without reliable data, attempts to correlate *in vivo* antitumor activity with clinical activity, which is the first aim of the project, produce doubtful results
- to perform multicenter *in vivo* experiments
- to critically debate tumor growth curves with complex biological patterns
- to use adaptive designs with interim stopping rules in preclinical research (refer to Table 1.2.5.1)
- to use historical data to improve statistical designs of *in vivo* antitumor activity experiments (refer to Table 1.2.5.1).

The first aim of the project was to correlate the quality of statistical design and analysis of preclinical *in vivo* experiments with estimates of clinical antitumor activity. In order to reach this aim, it would be necessary to identify a subgroup of *in vivo* experiments with good methodological quality. Instead, this survey showed that the quality of statistical design and analysis was systematically poor. Therefore, the first aim of the project failed. This aim of the project was probably too ambitious. Even if a subgroup of *in vivo* experiments was identified, the positive correlation between preclinical quality and clinical activity could not be detected for the following three reasons:

1. the effect size of the biological signal prevails over the methodological quality (i.e. a large effect size may be detected even in the presence of poor methodological quality)
2. only a very small part of the clinical effect is explained by the biological signal
3. the heterogeneity of the clinical effect due to the choice of the drug schedule, the selection of patients, and the random error is so large that it prevails over the biological signal.

Even if problems 1 to 3 did not exist, attempts to correlate *in vivo* antitumor activity with clinical activity will produce doubtful results without reliable data.

Another disappointing result of this survey was that for more than half of CCs tested as monotherapy in EOC clinical trials, no eligible preclinical *in vivo* experiment was identified. Due to

the dramatic lack of data from preclinical *in vivo* experiments, testing these CCs in clinical trials is not supported by robust preclinical evidence.

This project improved the methodology applied to tumor growth studies. An entirely innovative methodological framework to design and analyze preclinical *in vivo* tumor growth studies was proposed. This framework is complete because hypothesis tests are combined with coherent estimators of biological signal. Further studies comparing this framework with other statistical approaches are needed, but advantages of this new framework are evident:

- the statistical approach takes into account the complete dataset of *in vivo* tumor growth data
- the family of statistical tests is non-parametric. The parameter Δ_r , $r=0,\dots,K-1$, assume an additive effect at the time interval $[t_r, t_{r+1}]$. These parameters can vary between time intervals. There is a remarkable similarity between the framework proposed for the design and analysis of *in vivo* tumor growth studies and that framework usually used in survival analysis. The Mann-Whitney statistic and the family of statistical tests based on slopes computed from the experimental and control arms, corresponds to the Mantel-Haenzel statistic and the log-rank test, respectively. The parameter Δ_r , expressing the effect size as a relative change, corresponds to the parameter hazard ratio
- in case of missing values or right-censoring, data series of tumor growth curves are fully included
- if bootstrap methods and resampling techniques are used to compute the exact distribution, it would not be required to assume or specify a correlation structure. Otherwise the covariance matrix could be estimated by experimental data
- similar to the log-rank test, whose power is maximized if the proportional hazards assumption holds, the power of the proposed family of statistical tests is maximized under the alternative hypothesis of an additive effect
- the proposed framework could address every biological mechanism underlying tumor growth curves. Partitioning the follow-up in time intervals $[t_r, t_{r+1}]$, $r=0,\dots,K-1$, maximizes the biological information obtained from preclinical *in vivo* experiments
- because test statistic and estimators are calculated using the slopes that join the values of tumor volumes at times r and $r+1$, $r=0,\dots,K-1$, imbalances of tumor volumes at baseline are automatically adjusted. Moreover, because statistical tests are stratified by time interval $[r,$

$r+1]$, $r=0,\dots,K-1$, each time interval is an experimental block. The power of our statistical tests benefits from the blocked analysis

- patterns of tumor growth in the first days are well-balanced by patterns of tumor growth in the following days. Clearly, if tumor growth is reduced or just controlled by the experimental therapy in the first days, this advantage is formally lost if, in the following days, the pattern of tumor growth in the experimental arm is just the same as that in the control arm. In other words, the experimental therapy cannot maintain its initial advantage. This is important in the activity of screening CCs.

In conclusion,

- this project showed methodological limits and pitfalls of preclinical *in vivo* antitumor activity experiments performed in recent years. Preclinical researchers in Oncology should be aware of the limits and pitfalls shown in Tables 4.5.1-4.5.8, in order to improve statistical design, analysis and reporting of their preclinical *in vivo* experiments
- this project showed that the issue of fairly estimating and then debating the magnitude and pattern of the biological signal is poorly addressed by preclinical researchers, at least in Oncology. It is time to shift the researcher's interest from the mere presence of a treatment effect (i.e. the statistical significance is enough) to the estimation and judgement of the magnitude and pattern of the biological signal
- a new, useful and practical methodological framework to design, analyze and report *in vivo* tumor growth studies is proposed. It should be considered by preclinical researchers in Oncology for their *in vivo* antitumor activity experiments.

References

- Adjei AA, Christian M, Ivy P. Novel designs and end points for phase II clinical trials. *Clin Cancer Res*. 2009 Mar 15;15(6):1866-72
- Altman DG, De Stavola BL, Love SB, Stepniowska KA. Review of survival analyses published in cancer journals. *Br J Cancer*. 1995 Aug;72(2):511-8
- Bast RC Jr, Klug TL, St John E, Jenison E, Niloff JM, Lazarus H, Berkowitz RS, Leavitt T, Griffiths CT, Parker L, Zurawski VR Jr, Knapp RC. A radioimmunoassay using a monoclonal antibody to monitor the course of epithelial ovarian cancer. *N Engl J Med* 1983 Oct 13;309(15):883-887
- Bast RC, Thigpen JT, Arbuck SG, Basen-Engquist K, Burke LB, Freedman R, Horning SJ, Ozols R, Rustin GJ, Spriggs D, Wenzel LB, Pazdur R. Clinical trial endpoints in ovarian cancer: report of an FDA/ASCO/AACR Public Workshop. *Gynecol Oncol*. 2007 Nov;107(2):173-6
- Bebarta V, Luyten D, Heard K. Emergency medicine animal research: does use of randomization and blinding affect the results? *Acad Emerg Med*. 2003 Jun;10(6):684-7
- Begg AC. Analysis of growth delay data: potential pitfalls. *Br J Cancer Suppl*. 1980 Apr;4:93-7
- Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. *Nature*. 2012 Mar 28;483(7391):531-3
- Bell CB, Moser JM, Thompson R. Goodness Criteria for two-sample distribution-free tests. *Ann. Math. Statist*. 1966;37:133-142
- Bickel PJ. Edgeworth expansions in nonparametric statistics. *Ann. Statist*. 1974;2:1-20
- Boltze J, Wagner DC, Henninger N, Plesnila N, Ayata C. Phase III Preclinical Trials in Translational Stroke Research: Community Response on Framework and Guidelines. *Transl Stroke Res*. 2016 Aug;7(4):241-7
- Booth CM, Calvert AH, Giaccone G, Lobbezoo MW, Eisenhauer EA, Seymour LK. Design and conduct of phase II studies of targeted anticancer therapy: recommendations from the task force on methodology for the development of innovative cancer therapies (MDICT). *Eur J Cancer*. 2008 Jan;44(1):25-9
- Casella G. *Statistical Design*. 2008 Springer Science+Business Media
- Chamuleau SAJ, van der Naald M, Climent AM, Kraaijeveld AO, Wever KE, Duncker DJ, Fernández-Avilés F, Bolli R; Transnational Alliance for Regenerative Therapies in Cardiovascular Syndromes (TACTICS) Group. Translational Research in Cardiovascular Repair: A Call for a Paradigm Shift. *Circ Res*. 2018 Jan 19;122(2):310-318

- Corbett TH, White K, Polin L, Kushner J, Paluch J, Shih C, Grossman CS. Discovery and preclinical antitumor efficacy evaluations of LY32262 and LY33169. *Invest New Drugs*. 2003 Feb;21(1):33-45
- De Angelis C, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, Kotzin S, Laine C, Marusic A, Overbeke AJ, Schroeder TV, Sox HC, Van Der Weyden MB; International Committee of Medical Journal Editors. Clinical trial registration: a statement from the International Committee of Medical Journal Editors. *N Engl J Med*. 2004 Sep 16;351(12):1250-1
- Demidenko E. *Mixed Model: Theory and Applications*. 2004 Wiley-Interscience
- Demidenko E. The Assessment of Tumour Response to Treatment. *Journal of the Royal Statistical Society. Series C*. Vol. 55, No. 3 (2006), 365-377
- DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. *J Health Econ*. 2016 May;47:20-33
- Dirnagl U, Fisher M. International, multicenter randomized preclinical trials in translational stroke research: it's time to act. *J Cereb Blood Flow Metab*. 2012 Jun;32(6):933-5
- Eisenhart C. The meaning of 'least' in least squares. *J. Wash. Acad. Sci*. 1964;54:24-33
- Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, Dancey J, Arbuck S, Gwyther S, Mooney M, Rubinstein L, Shankar L, Dodd L, Kaplan R, Lacombe D, Verweij J. New response evaluation criteria in solid tumors: Revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009 Jan;45(2):228-47
- Festing MF. Statistics and animals in biomedical research. *Significance*. 2010;7:176 – 177
- Festing MF. Randomized block experimental designs can increase the power and reproducibility of laboratory animal experiments. *ILAR J*. 2014;55(3):472-6
- Festing MF. On determining sample size in experiments involving laboratory animals *Lab Anim*. 2018 Aug;52(4):341-350
- Festing MF, Altman DG. Guidelines for the design and statistical analysis of experiments using laboratory animals. *ILAR J*. 2002;43(4):244-58
- Festing MF, Baumans V, Combes RD, Halder M, Hendriksen CF, Howard BR, Lovell DP, Moore GJ, Overend P, Wilson MS. Reducing the use of laboratory animals in biomedical research: problems and possible solutions. *Altern Lab Anim*. 1998 May-Jun;26(3):283-301
- Fine JP, Gray RJ. A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*. 1999;94:496-509
- Fisher RA. *The Design of Experiments*, chapter 2. 1935 Edinburgh: Oliver & Boyd

- Fitts DA., Ethics and animal numbers: informal analyses, uncertain sample sizes, inefficient replications, and type I errors. *J Am Assoc Lab Anim Sci*. 2011 Jul;50(4):445-53
- Fix E, Hodges JL. Significance probabilities for the Wilcoxon Test. *Ann. Math. Statist*. 1955;26:301-312
- Frapolli R, Bello E, Ponzio M, Craparotta I, Mannarino L, Ballabio S, Marchini S, Carrassa L, Ubezio P, Porcu L, Brich S, Sanfilippo R, Casali PG, Gronchi A, Pilotti S, D'Incalci M. Combination of PPAR γ agonist pioglitazone and trabectedin induce adipocyte differentiation to overcome trabectedin resistance in myxoid liposarcomas. *Clin Cancer Res*. 2019 Sep 3. pii: clincanres.0976.2019
- Garattini S, Grignaschi G. Animal testing is still the best way to find new treatments for patients. *Eur J Intern Med*. 2017 Apr;39:32-35
- Gehan EA. A generalized Wilcoxon test for comparing arbitrarily singly censored samples. *Biometrika*. 1965 Jun;52: 203-223
- Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis*. Second Edition. 2004 Chapman & Hall/CRC
- Gladstone DJ, Black SE, Hakim AM. Toward wisdom from failure: lessons from neuroprotective stroke trials and new therapeutic directions. *Stroke*. 2002 Aug;33(8):2123-36
- Goodman SN, Fanelli D, Ioannidis JP. What does research reproducibility mean? *Sci Transl Med*. 2016 Jun 1;8(341):341ps12
- Gray RJ. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Statistics*. 1988;16: 1141-1154
- Green S, Weiss GR. Southwest Oncology Group standard response criteria, endpoint definitions and toxicity criteria. *Invest New Drugs*. 1992 Nov;10(4):239-53
- Grotta J. Neuroprotection is unlikely to be effective in humans using current trial designs. *Stroke*. 2002 Jan;33(1):306-7
- Hall P. (1992) Principles of Edgeworth Expansion. In: *The Bootstrap and Edgeworth Expansion*. Springer Series in Statistics. Springer, New York, NY
- Hanahan D, Weinberg RA. The hallmarks of cancer. *Cell*. 2000 Jan 7;100(1):57-70
- Harris R, Bradburn M, Jon Deeks J, Harbord Roger, Altman DG, Steichen T, Sterne J, 2006. "METAN: Stata module for fixed and random effects meta-analysis," Statistical Software Components S456798, Boston College Department of Economics, revised 23 Sep 2010
- Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. Clinical development success rates for investigational drugs. *Nat Biotechnol*. 2014 Jan;32(1):40-51

- Haynam GE, Govindarajulu Z. Exact power of Mann-Whitney tests for exponential and rectangular alternatives. *Ann. Math. Statist.* 1966;37:945-953
- Heitjan DF, Manni A, Santen RJ. Statistical analysis of in vivo tumor growth experiments. *Cancer Res.* 1993 Dec 15;53(24):6042-50
- Henderson VC, Demko N, Hakala A, MacKinnon N, Federico CA, Fergusson D, Kimmelman J. A meta-analysis of threats to valid clinical inference in preclinical research of sunitinib. *Elife.* 2015 Oct 13;4:e08351
- Hendriks HR, Langdon S, Berger DP, Breistøl K, Fiebig HH, Fodstad O, Schwartzmann G. Comparative antitumour activity of vinblastine-isoleucinate and related vinca alkaloids in human tumour xenografts. *Eur J Cancer.* 1992;28A(4-5):767-73
- Hooijmans CR, Rovers MM, de Vries RB, Leenaars M, Ritskes-Hoitinga M, Langendam MW. SYRCLE's risk of bias tool for animal studies *BMC Medical Research Methodology* 2014, 14:43
- Houghton PJ, Morton CL, Tucker C, Payne D, Favours E, Cole C, Gorlick R, Kolb EA, Zhang W, Lock R, Carol H, Tajbakhsh M, Reynolds CP, Maris JM, Courtright J, Keir ST, Friedman HS, Stopford C, Zeidner J, Wu J, Liu T, Billups CA, Khan J, Ansher S, Zhang J, Smith MA. The pediatric preclinical testing program: description of models and early testing results. *Pediatr Blood Cancer.* 2007 Dec;49(7):928-40
- Jüni P, Altman DG, Egger M. Systematic reviews in health care: Assessing the quality of controlled clinical trials. *BMJ.* 2001 Jul 7;323(7303):42-6
- Kalbfleish JD, Prentice RL. *The Statistical Analysis of Failure Time Data.* 1980 New York: John Wiley & Sons, Inc
- Kernan WN, Viscoli CM, Makuch RW, Brass LM, Horwitz RI. Stratified randomization for clinical trials. *J Clin Epidemiol.* 1999;52(1):19-26
- Kilkenney C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: The ARRIVE guidelines for reporting animal research. *PLoS Biol.* 2010 Jun 29;8(6):e1000412
- Kilkenney C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, Fry D, Hutton J, Altman DG. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLoS One.* 2009 Nov 30;4(11):e7824
- Klotz J. The Wilcoxon. Ties and the computer. *J. Am. Statist. Assoc.* 1966;61:772-787
- Langley GR, Adcock IM, Busquet F, Crofton KM, Csernok E, Giese C, Heinonen T, Herrmann K, Hofmann-Apitius M, Landesmann B, Marshall LJ, McIvor E, Muotri AR, Noor F, Schutte K, Seidle

- T, van de Stolpe A, Van Esch H, Willett C, Woszczek G. Towards a 21st-century roadmap for biomedical research and drug discovery: consensus report and recommendations. *Drug Discov Today*. 2017 Feb;22(2):327-339
- Lehmann EL. *Nonparametrics Statistical Methods Based on Ranks*. Revised First Edition. 2006 Springer Science+Business Media
 - Lehmann EL, Casella G. *Theory of Point Estimation*. Second Edition. 1998 Springer-Verlag
 - Lehmann EL, Romano JP. *Testing Statistical Hypotheses*. Third Edition. 2005 Springer Science+Business Media
 - Liang H. Modeling antitumor activity in xenograft tumor treatment. *Biom J*. 2005 Jun;47(3):358-68
 - Liang H. Comparison of antitumor activities in tumor xenograft treatment. *Contemp Clin Trials*. 2007 Feb;28(2):115-9
 - Liang H, Sha N. Modeling antitumor activity by using a non-linear mixed-effects model. *Math Biosci*. 2004 May;189(1):61-73
 - Maertens O, McCurrach ME, Braun BS, De Raedt T, Epstein I, Huang TQ, Lauchle JO, Lee H, Wu J, Cripe TP, Clapp DW, Ratner N, Shannon K, Cichowski K. A Collaborative Model for Accelerating the Discovery and Translation of Cancer Therapies. *Cancer Res*. 2017 Nov 1;77(21):5706-5711
 - Michaelis LC, Ratain MJ. Measuring response in a post-RECIST world: from black and white to shades of grey. *Nat Rev Cancer*. 2006 May;6(5):409-14
 - Milton RC. *Rank order probabilities: two-sample normal shift alternatives*. 1970 John Wiley & Sons, New York
 - Morgan S, Grootendorst P, Lexchin J, Cunningham C, Greyson D. The cost of drug development: a systematic review. *Health Policy*. 2011 Apr;100(1):4-17
 - Nyaga NV, Arbyn M, Aerts M, 2014. "METAPROP: Stata module to perform fixed and random effects meta-analysis of proportions," *Statistical Software Components S457781*, Boston College Department of Economics, revised 26 Jan 2017
 - O'Brien TJ, Ben-Menachem E, Bertram EH 3rd, Collins SD, Kokaia M, Lerche H, Klitgaard H, Staley KJ, Vaudano E, Walker MC, Simonato M. Proposal for a "phase II" multicenter trial model for preclinical new antiepilepsy therapy development. *Epilepsia*. 2013 Aug;54 Suppl 4:70-4
 - Ocana A, Pandiella A, Siu LL, Tannock IF. Preclinical development of molecular-targeted agents for cancer. *Nat Rev Clin Oncol*. 2010 Dec 7;8(4):200-9

- Oken M, Creech R, Tormey D, et al. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *Am J Clin Oncol*. 1982;5:649-655
- Pearl ML, Yashar CM, Johnston CM, Reynolds RK, Roberts JA. Exponential regression of CA 125 during salvage treatment of ovarian cancer with taxol. *Gynecol Oncol* 1994;53:339–43
- Prentice RL. Surrogate endpoints in clinical trials: definition and operational criteria. *Stat Med*. 1989 Apr;8(4):431-40
- Prinz F, Schlange T, Asadullah K. Believe it or not: how much can we rely on published data on potential drug targets? *Nat Rev Drug Discov*. 2011 Aug 31;10(9):712
- Ratain MJ. Phase II oncology trials: let's be positive. *Clin Cancer Res*. 2005 Aug 15;11(16):5661-2
- Ratain MJ, Sargent DJ. Optimising the design of phase II oncology trials: the importance of randomization. *Eur J Cancer*. 2009 Jan;45(2):275-80
- Rick Ng. *Drugs: From discovery to approval*. 3rd ed. New Jersey: Wiley-Blackwell; 2015
- Rubin SC. *Chemotherapy of Gynecologic Cancers: Society of Gynecologic Oncologists Handbook*. Philadelphia: Lippincott Williams & Wilkins; 2004
- Russell WMS and Burch RL. *The Principles of Humane Experimental Technique*. 1992 Special edition. Potters Bar: Universities Federation for Animal Welfare
- Rustin GJ, Bast RC Jr, Kelloff GJ, Barrett JC, Carter SK, Nisen PD, Sigman CC, Parkinson DR, Ruddon RW. Use of CA-125 in clinical trial evaluation of new therapeutic drugs for ovarian cancer. *Clin Cancer Res*. 2004 Jun 1;10(11):3919-26
- Rustin GJ, Nelstrop AE, McClean P, Brady MF, McGuire WP, Hoskins WJ, Mitchell H, Lambert HE. Defining response of ovarian carcinoma to initial chemotherapy according to serum CA 125. *J Clin Oncol*. 1996 May;14(5):1545-51
- Rustin GJ, Timmers P, Nelstrop A, Shreeves G, Bentzen SM, Baron B, Piccart MJ, Bertelsen K, Stuart G, Cassidy J, Eisenhauer E. Comparison of CA-125 and standard definitions of progression of ovarian cancer in the Intergroup trial of cisplatin and paclitaxel versus cisplatin and cyclophosphamide. *J Clin Oncol*. 2006 Jan 1;24(1):45-51
- Sena ES, van der Worp HB, Bath PM, Howells DW, Macleod MR. Publication bias in reports of animal stroke studies leads to major overstatement of efficacy. *PLoS Biol*. 2010 Mar 30;8(3):e1000344
- Seymour L, Ivy SP, Sargent D, Spriggs D, Baker L, Rubinstein L, Ratain MJ, Le Blanc M, Stewart D, Crowley J, Groshen S, Humphrey JS, West P, Berry D. The design of phase II clinical trials testing cancer therapeutics: consensus recommendations from the clinical trial design task force of the

national cancer institute investigational drug steering committee Clin Cancer Res. 2010 Mar 15;16(6):1764-9

- Sharma MR, Maitland ML, Ratain MJ. RECIST: no longer the sharpest tool in the oncology clinical trials toolbox---point. Cancer Res. 2012 Oct 15;72(20):5145-9; discussion 5150
- Schoenfeld D. The asymptotic properties of nonparametric tests for comparing survival distributions. Biometrika. 1981 Apr;68:316-319
- Schoenfeld DA. Sample-Size Formula for the Proportional-Hazards Regression Model. Biometrics. 1983 Jun;39:499-503
- Simpson T. A letter to the Right Honorable George Earl of Macclesfield, President of the Royal Society, on the advantage of taking the mean of a number of observations, in practical astronomy. Phil. Trans. R. Soc. London. 1755;49(Pt.1):82-93
- Smith AD, Shah SN, Rini BI, Lieber ML, Remer EM. Morphology, Attenuation, Size, and Structure (MASS) criteria: assessing response and predicting clinical outcome in metastatic renal cell carcinoma on antiangiogenic targeted therapy. AJR Am J Roentgenol. 2010 Jun;194(6):1470-8
- Stadler W. New trial designs to assess antitumor and antiproliferative agents in prostate cancer. Invest New Drugs. 2002 May;20(2):201-8
- Szakács G, Gottesman MM. Comparing solid tumors with cell lines: implications for identifying drug resistance genes in cancer. Mol Interv. 2004 Dec;4(6):323-5
- Tan M, Fang HB, Tian GL, Houghton PJ. Small-sample inference for incomplete longitudinal data with truncation and censoring in tumor xenograft models. Biometrics. 2002 Sep;58(3):612-20
- Tarone RE, Ware J. On distribution free tests for equality of survival distributions. Biometrika. 1977;64:156-160
- Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, van Oosterom AT, Christian MC, Gwyther SG. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst. 2000 Feb 2;92(3):205-16
- Thiesse P, Ollivier L, Di Stefano-Louineau D, Négrier S, Savary J, Pignard K, Lasset C, Escudier B. Response rate accuracy in oncology trials: reasons for interobserver variability. Groupe Français d'Immunothérapie of the Fédération Nationale des Centres de Lutte Contre le Cancer. J Clin Oncol. 1997 Dec;15(12):3507-14

- Tonkin K, Tritchler D, Tannock I. Criteria of tumor response used in clinical trials of chemotherapy. *J Clin Oncol*. 1985 Jun;3(6):870-5
- Ubezio P. Beyond The T/C Ratio: Old And New Anticancer Activity Scores In Vivo. *Cancer Management and Research* 2019; 11:8529–8538
- van der Burg ME, Lammes FB, Verweij J. The role of CA 125 and conventional examinations in diagnosing progressive carcinoma of the ovary. *Surg Gynecol Obstet*. 1993 Apr;176(4):310-4
- van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, Macleod MR. Can animal models of disease reliably inform human studies? *PLoS Med*. 2010 Mar 30;7(3):e1000245
- Vardi Y, Zhiliang YZ, Zhang CH. Two-Sample Tests for Growth Curves under Dependent Right Censoring. *Biometrika* Vol. 88, No. 4 (Dec., 2001), 949-960
- Verdooren LR. Extended tables of critical values for Wilcoxon's test statistic. *Biometrika*. 1963;50:177-186
- Vergote I, Rustin GJ, Eisenhauer EA, Kristensen GB, Pujade-Lauraine E, Parmar MK, Friedlander M, Jakobsen A, Vermorken JB (2000) Re: new guidelines to evaluate the response to treatment in solid tumors [ovarian cancer] Gynecologic Cancer Intergroup. *J Natl Cancer Inst* 92: 1534– 1535
- Warr D, McKinney S, Tannock I. Influence of measurement error on assessment of response to anticancer chemotherapy: proposal for new criteria of tumor response. *J Clin Oncol*. 1984 Sep;2(9):1040-6
- Warr D, McKinney S, Tannock I. Influence of measurement error on response rates. *Cancer Treat Rep*. 1985 Oct;69(10):1127-32
- Wedge SR, Kendrew J, Hennequin LF, Valentine PJ, Barry ST, Brave SR, Smith NR, James NH, Dukes M, Curwen JO, Chester R, Jackson JA, Boffey SJ, Kilburn LL, Barnett S, Richmond GH, Wadsworth PF, Walker M, Bigley AL, Taylor ST, Cooper L, Beck S, Jürgensmeier JM, Ogilvie DJ. AZD2171: a highly potent, orally bioavailable, vascular endothelial growth factor receptor-2 tyrosine kinase inhibitor for the treatment of cancer. *Cancer Res*. 2005 May 15;65(10):4389-400
- Williams S. Surrogate endpoints in early prostate cancer research *Transl Androl Urol*. 2018 Jun;7(3):472-482
- Wu J, Houghton PJ. Interval approach to assessing antitumor activity for tumor xenograft studies. *Pharm Stat*. 2010 Jan-Mar;9(1):46-54
- Zhao L, Morgan MA, Parsels LA, Maybaum J, Lawrence TS, Normolle D. Bayesian hierarchical changepoint methods in modeling the tumor growth profiles in xenograft experiments. *Clin Cancer Res*. 2011 Mar 1;17(5):1057-64

- Zitvogel L, Pitt JM, Daillère R, Smyth MJ, Kroemer G. Mouse models in oncoimmunology. *Nat Rev Cancer*. 2016 Dec;16(12):759-773

Appendix A

A.1 Preclinical search string used in the Medline database

Search	Query	N° items
#58	#1 AND #2 AND #3 AND #57	613
#57	#4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16 OR #17 OR #18 OR #19 OR #20 OR #21 OR #22 OR #23 OR #24 OR #25 OR #26 OR #27 OR #28 OR #29 OR #30 OR #31 OR #32 OR #33 OR #34 OR #35 OR #36 OR #37 OR #38 OR #39 OR #40 OR #41 OR #42 OR #43 OR #44 OR #45 OR #46 OR #47 OR #48 OR #49 OR #50 OR #51 OR #52 OR #53 OR #54 OR #55 OR #56	169651
#56	("Paclitaxel"[Mesh] OR paclitaxel OR Abraxane OR ABI007 OR ABI-007 OR "ABI 007" OR abi007 OR abraxane OR anzatax OR "bms 181339" OR endotag-1 OR genexol OR "genexol pm" OR infinnium OR intaxel OR "mbt 0206" OR "mitotax" OR "nab paclitaxel" OR "nanoparticle albumin bound paclitaxel" OR "nsc 125973" OR nsc125973 OR "oncogel" OR "onxol" OR "pacitaxel" OR "paclitaxel nab" OR padexol OR parexel OR paxceed OR "paxene" OR "paxus" OR "praxel" OR taxol OR yewtaxan) AND ("lipid nanoparticles" OR "lipid core nanoparticles")	92
#55	"gemcitabine" [Supplementary Concept] OR gemcitabine OR dFdCyd OR "2',2'-difluorodeoxycytidine" OR "2'-deoxy-2'-difluorocytidine" OR "2',2'-difluoro-2'- deoxycytidine" OR "LY 188011" OR LY-188011 OR Gemzar OR "2' deoxy 2', 2' Difluorocytidine" OR "2', 2' Difluorodeoxycytidine" OR d07001 OR difluorodeoxycytidine OR gemcite OR gemzar OR "ly 188011" OR ly188011	15711
#54	"Lenalidomide"[Mesh] OR lenalidomide OR "3-(4-Amino-1-oxoisindolin- 2-yl)piperidine-2,6-dione" OR "CC 5013" OR CC5013 OR CC-5013 OR Revlimid OR "Revimid" OR "cc 5013" OR cc5013 OR "cdc 501" OR "enmd 0997" OR "imid 3" OR imid3 OR "revimid" OR revlimid	4114
#53	"birinapant" [Supplementary Concept] OR birinapant OR tl32711	68
#52	"belinostat" [Supplementary Concept] OR belinostat OR beleodaq OR pdx101 OR "pxd 101"	231
#51	"nintedanib" [Supplementary Concept] OR "nintedanib" OR "bibf 1120" OR bibf1120 OR intedanib OR ofev OR vargatef OR "BIBF 1120" OR BIBF1120 OR BIBF-1120	716
#50	"AZD 6244" [Supplementary Concept] OR "AZD 6244" OR AZD6244 OR AZD-6244 OR selumetinib OR "ARRY 142886" OR ARRY142886 OR ARRY-142886 OR "arry 142886" OR arry142886 OR "azd 6244" OR azd6244	563
#49	"aflibercept" [Supplementary Concept] OR aflibercept OR "VEGF Trap - regeneron" OR "VEGF Trap-Eye" OR VEGF-Trap OR eylea OR Zaltrap OR "AVE 0005" OR AVE0005 OR AVE-0005 OR "AVE 005" OR AVE005 OR AVE-005 OR ZIV-aflibercept OR "vascular endothelial growth factor trap" OR "VEGF Trap" OR "ziv aflibercept"	1953
#48	"N-(4-bromo-2-fluorophenyl)-6-methoxy-7-((1-methylpiperidin- 4-yl)methoxy)quinazolin-4-amine" [Supplementary Concept] OR vandetanib OR "N-(4- bromo-2-fluorophenyl)-6-methoxy-7-((1-methylpiperidin-4-yl)methoxy)quinazolin- 4-amine" OR "caprelsa" OR Zactima OR "ZD 6474" OR ZD6474 OR ZD-6474 OR "caprelsa" OR "vandetinib" OR zactima OR "zd 6474" OR zd6474	874
#47	"prexasertib" [Supplementary Concept] OR prexasertib OR LY2606368 OR ly2606368	24
#46	"Trabectedin"[Mesh] OR Trabectedin OR Yondelis OR "Ecteinascidin 743" OR ET-743 OR ET743 OR "ET 743" OR "NSC 684766" OR "ecteinascidin 743" OR "et 743" OR et743 OR yondelis	827
#45	"ST 1481" [Supplementary Concept] OR "ST 1481" OR ST1481 OR LBQ707 OR 7-t-butoxyiminomethylcamptothecin OR 7-t-butoxyiminomethyl-camptothecin OR gimatecan OR "7 tert butoxyiminomethylcamptothecin" OR lbq707 OR "st 1481" OR st1481	52

#44	"Dasatinib"[Mesh] OR dasatinib OR Sprycel OR "BMS 354825" OR BMS354825 OR BMS-354825 OR "bms 354825" OR bms354825 OR "n (2 chloro 6 methylphenyl) 2 [[6 [4 (2 hydroxyethyl) piperazin 1 yl] 2 methylpyrimidin 4 yl] amino] thiazole 5 carboxamide" OR sprycel	3193
#43	"ixabepilone" [Supplementary Concept] OR ixabepilone OR "azaepothilone B" OR BMS247550 OR "BMS 247550" OR BMS-247550 OR "azaepothilone B" OR "bms 247550" OR bms247550 OR ixempra	426
#42	"perifosine" [Supplementary Concept] OR perifosine OR "octadecyl-(1,1- dimethyl-4-piperidylo)phosphate" OR "D 21266" OR D-21266 OR "d 21266" OR d21266 OR "krx 0401"	353
#41	"apatinib" [Supplementary Concept] OR rivoceranib OR apatinib OR YN968D1 OR YN-968D1 OR "apatinib mesylate" OR "aitan" OR apatinib OR "apatinib mesylate" OR yn968d1	311
#40	"BI 6727" [Supplementary Concept] OR BI-6727 OR volasertib OR "bi 6727" OR bi6727	121
#39	"temsirolimus" [Supplementary Concept] OR temsirolimus OR "rapamycin, 42-(3-hydroxy-2-(hydroxymethyl)-2-methylpropanoate)" OR "CCI 779" OR CCI-779 OR Torisel OR "cci 779" OR cci779 OR "cell cycle inhibitor 779" OR "nsc 683864" OR "rapamycin 42 [2, 2 bis (hydroxymethyl) propionate]" OR torisel OR "way-cci 779"	1525
#38	"Imatinib Mesylate"[Mesh] OR imatinib OR "Imatinib Methanesulfonate" OR STI571 OR STI-571 OR "STI 571" OR Gleevec OR Glivec OR "ST 1571" OR ST1571 OR "CGP 57148" OR CGP57148B OR CGP-57148 OR CGP57148 OR Imatinib OR "Alpha-(4- methyl-1-piperazinyl)-3'-((4-(3-pyridyl)-2-pyrimidinyl)amino)-p-tolu-p-toluidide" OR "cgp 57148" OR "cgp 57148b" OR cgp57148 OR cgp57148b OR "gleevac" OR gleevec OR glivec OR "glivic" OR "imatinib mesilate" OR "imatinib mesylate" OR "signal transduction inhibitor 571" OR "st 1571" OR st1571 OR "sti 571" OR sti-571 OR sti571	14892
#37	"MLN 8237" [Supplementary Concept] OR "MLN 8237" OR MLN8237 OR MLN-8237 OR alisertib OR "mln 8237" OR mln8237	271
#36	"SPI-77, liposomal" [Supplementary Concept] OR "liposomal cisplatin" OR SPI-077 OR "SPI 077" OR "SPI 77" OR SPI-77 OR SPI077 OR "Stealth liposomal cisplatin" OR "cisplatin liposomal" OR (cisplatin AND (liposomes OR liposomal))	544
#35	"PM 01183" [Supplementary Concept] OR "PM 01183" OR PM01183 OR PM-01183 OR lurbinedectin OR "pm 01183" OR pm01183	39
#34	"ALK1-Fc fusion protein, human" [Supplementary Concept] OR "ALK1-Fc fusion protein, human" OR ACE-041 OR dalantercept OR "ace 041"	14
#33	"olaparib" [Supplementary Concept] OR olaparib OR "AZD 2281" OR AZD2281 OR AZD-2281 OR AZD221 OR Lynparza OR "4 [3 (4 cyclopropanecarbonylpiperazine 1 carbonyl) 4 fluorobenzyl] 2h phthalazin 1 one" OR "4 [[3 [[4 (cyclopropylcarbonyl) piperazin 1 yl] carbonyl] 4 fluorophenyl] methyl] phthalazin 1 (2h) one" OR "azd 2281" OR azd2281 OR "ku 0059436" OR "ku 59436" OR ku0059436 OR lynparza	1035
#32	"Doxorubicin"[Mesh] OR doxorubicin OR Farmiblastina OR "Rubex" OR Adriamycin OR Adriblastin OR "Adriblastine" OR Adriblastina OR "Adriablastine" OR Adriablastin OR DOXO-cell OR "DOXO cell" OR "Myocet" OR "14 hydroxydaunomycin" OR adriablastin OR "adriablastina" OR "adriablastine" OR "adriacin" OR "adriamicina" OR "adriamicine" OR adriamycin OR "adriamycin hydrochloride" OR "adriamycin rdf" OR "adriamycina" OR adriblastin OR adriblastina OR "adriblastine" OR "adrim" OR "caelix" OR caelyx OR "caelyx/doxil" OR "doxorubicin" OR "dox sl" OR dox-sl OR doxil OR "doxil (liposomal)" OR "doxorubicin hydrochloride" OR "doxorubicin, liposomal" OR "doxorubicine" OR "doxorubin" OR "evacet" OR farmiblastina OR "fi 106" OR lipodox OR "liposomal doxorubicin" OR "mcc 465" OR mcc465 OR "myocet" OR "nsc 123127" OR "pegylated liposomal doxorubicin" OR "polyethylene glycol-coated liposomal doxorubicin" OR "rubex" OR "tlc d 99"	73070
#31	"BI 2536" [Supplementary Concept] OR "BI 2536" OR BI2536 OR BI-2536	202
#30	"Bendamustine Hydrochloride"[Mesh] OR Bendamustine OR "Bendamustin" OR Cytostasan OR "IMET 3393" OR Ribomustin OR Treanda OR "Zimet 3393" OR "5 [bis (2 chloroethyl) amino] 1	1186

	methyl 2 benzimidazolebutyric acid" OR "bendamustine hydrochloride" OR "bendeka" OR cytosasan OR "cytosasane" OR "imet 3393" OR "levact" OR ribomustin OR treanda OR "zimet 3393"	
#29	"eribulin" [Supplementary Concept] OR eribulin OR "E 7389" OR E-7389 OR ER-086526 OR ER086526 OR "ER 086526" OR ER-86526 OR Halaven OR "NSC 707389" OR NSC707389 OR NSC-707389 OR "B 1793" OR B-1793 OR "B 1939" OR B-1939 OR "eribulin mesylate" OR "eribulin mesilate" OR "e 7389" OR e7389 OR "er 086526" OR er086526 OR "eribulin mesilate" OR "eribulin mesylate" OR halaven	594
#28	"danusertib" [Supplementary Concept] OR danusertib OR "PHA 739358" OR PHA739358 OR PHA-739358 OR "pha 739358" OR pha739358	63
#27	"LHRH, lysine(6)-doxorubicin" [Supplementary Concept] OR "LHRH, lysine(6)- doxorubicin" OR "zoptarelin doxorubicin" OR ZEN-008 OR AN-152 OR AEZS-108 OR "aezs 108" OR "an 152" OR an-152	62
#26	"Sorafenib"[Mesh] OR sorafenib OR Nexavar OR "BAY 43-9006" OR "BAY 43 9006" OR "BAY 439006" OR "Sorafenib N-Oxide" OR "Sorafenib N Oxide" OR BAY-673472 OR BAY-545-9085 OR BAY5459085 OR "Sorafenib Tosylate" OR "bay 43 9006" OR "bay 43-9006" OR "bay 439006" OR "bay43 9006" OR bay43-9006 OR bay439006 OR nexavar OR "sorafenib tosylate"	7790
#25	"cabozantinib" [Supplementary Concept] OR cabozantinib OR Cometriq OR "XL 184" OR XL-184 OR "BMS 907351" OR BMS907351 OR BMS-907351 OR "bms 907351" OR bms907351 OR cabometyx OR "cabozantinib malate" OR "cabozantinib s malate" OR "cabozantinib s-malate" OR cometriq OR "xl 184" OR xl184	684
#24	"pazopanib" [Supplementary Concept] OR pazopanib OR GW786034B OR GW-786034B OR GW780604 OR GW-780604 OR Votrient OR "armala" OR "gw 786034" OR gw786034 OR gw786034b OR "pazopanib hydrochloride" OR votrient	1542
#23	"sagopilone" [Supplementary Concept] OR sagopilone OR DE-03757 OR EPO-477 OR SH-Y-03757 OR ZK-epothilone OR SH-Y03757A OR ZK-219477 OR ZKEPO OR BAY-86-5302 OR "zk 219477" OR "zk epo" OR zk219477	58
#22	"Topotecan"[Mesh] OR topotecan OR 9-Dimethylaminomethyl-10- hydroxycamptothecin OR "9 Dimethylaminomethyl 10 hydroxycamptothecin" OR "Topotecan Hydrochloride" OR "Nogitecan Hydrochloride" OR SKF-104864-A OR "SKF 104864 A" OR SKF104864A OR Hycamtin OR NSC-609699 OR "NSC 609699" OR NSC609699 OR "9 dimethylaminomethyl 10 hydroxycamptothecin" OR e89001 OR hycamptamine OR hycamtin OR "nsc 609699" OR nsc609699 OR "skf 104864" OR "skf 104864 a" OR "topotecan hydrochloride" OR "topotecane"	3178
#21	("2,2-dimethyl-N-(6-oxo-6,7-dihydro-5H-dibenzo(b,d)azepin-7-yl)-N'-(2,2,3,3,3-pentafluoropropyl)malonamide" [Supplementary Concept] OR RO4929097 OR RO-4929097)	46
#20	"epothilone B" [Supplementary Concept] OR "epothilone B" OR patupilone OR EPO906 OR "epo 906" OR epo906	458
#19	"acetyl-lysyl-prolyl-seryl-seryl-prolyl-prolyl-glutamyl-glutamic acid amide" [Supplementary Concept] OR ("urokinase plasminogen activator" AND "136-143") OR "urokinase derived peptide" OR "urokinase-derived peptide" OR ("A6" AND peptide) OR Ac-Lys-Pro-Ser-Ser-Pro-Pro-Glu-Glu-NH2 OR Ac-KPSSPPEE-NH2 OR "acetyl- lysylprolyl- seryl-seryl-prolyl-prolyl-glutamyl- glutamic acid amide"	906
#18	"tasquinimod" [Supplementary Concept] OR tasquinimod OR ABR-215050 OR "abr 215050"	64
#17	"veliparib" [Supplementary Concept] OR veliparib OR "2-((R)- 2-methylpyrrolidin-2-yl)-1H-benzimidazole-4-carboxamide" OR "ABT 888" OR ABT888 OR ABT-888 OR "abt 888" OR abt888	338
#16	"Sunitinib"[Mesh] OR sunitinib OR "Sunitinib Malate" OR Sutent OR "SU 11248" OR SU011248 OR "SU 011248" OR SU-011248 OR SU11248 OR SU-11248 OR "su 011248" OR "su 11248" OR su011248 OR su11248 OR "sunitinib malate" OR suo11248 OR sutent	5732

#15	"Albumin-Bound Paclitaxel"[Mesh]OR 130-nm albumin-bound paclitaxel [Supplementary Concept] OR (albumin AND paclitaxel) OR "130-nm albumin-bound paclitaxel" OR "Albumin-Bound Paclitaxel" OR nab-paclitaxel OR "nab paclitaxel" OR "Albumin Bound Paclitaxel" OR "Protein-Bound Paclitaxel" OR "Protein Bound Paclitaxel" OR Abraxane OR ABI007 OR ABI-007 OR "ABI 007" OR abi007 OR abraxane OR anzatax OR "bms 181339" OR endotag-1 OR genexol OR "genexol pm" OR infinnium OR intaxel OR "mbt 0206" OR "mitotax" OR "nab paclitaxel" OR "nanoparticle albumin bound paclitaxel" OR "nsc 125973" OR nsc125973 OR "oncogel" OR "onxol" OR "pacitaxel" OR "paclitaxel nab" OR padexol OR parexel OR paxceed OR "paxene" OR "paxus" OR "praxel" OR taxol OR yewtaxan	37435
#14	"motesanib diphosphate" [Supplementary Concept] OR motesanib OR "AMG 706" OR AMG706 OR AMG-706 OR "amg 706" OR amg706 OR "motesanib diphosphate"	97
#13	"Lapatinib"[Mesh] OR lapatinib OR Tykerb OR GW282974X OR GW-282974X OR GW572016 OR GW-572016 OR "GW 572016" OR "Lapatinib Ditosylate" OR "gw 572016" OR gw2016 OR gw572016 OR "lapatinib ditosylate" OR tykerb OR "tyverb"	2498
#12	"Asparaginase"[Mesh] OR "monomethoxypolyethylene glycol-conjugated asparaginase" [Supplementary Concept] OR asparaginase OR "monomethoxypolyethylene glycol-conjugated asparaginase" OR "L asparaginase" OR L-asparaginase OR "PEG(2)-ASP" OR "asparaginase 2" OR "asparaginase a" OR "asparaginase ag" OR "asparaginase Erwinia chrysanthemi" OR "asparaginase ii" OR "asparagine amidohydrolase" OR "asparginase" OR "aspariginase" OR colaspase OR crasnitin OR crisantaspase OR "e.c. 3.5.1.1" OR elspar OR erwinase OR "erwinaze" OR "ery asp" OR "graspa" OR kidrolase OR krasniti OR "I asparaginase" OR "I asparaginase a" OR "I asparagine amidohydrolase" OR "I asparginase" OR leunase OR "levo asparaginase" OR "nsc 109229" OR "Asparagine Deaminase" OR "Asparaginase II" OR Leunase OR "Asparaginase medac"	5615
#11	"rucaparib" [Supplementary Concept] OR rucaparib OR "AG 014699" OR AG014699 OR AG-014699 OR PF-01367338 OR "8 fluoro 2 [4 [(methylamino) methyl] phenyl] 1, 3, 4, 5 tetrahydro 6h azepino [5, 4, 3 cd] indol 6 one" OR "ag 014699" OR "ag 14447" OR "ag 14699" OR ag014699 OR ag14447 OR "co 338" OR "pf 01367338" OR "pf 1367338" OR rubraca	225
#10	("5'-oleoyl cytarabine" [Supplementary Concept] OR "5'-oleoyl cytarabine" OR "5'-oleyl-ara-C" OR elacyt OR elacytarabine OR "CP 4055" OR CP4055 OR CP-4055)	33
#9	"iniparib" [Supplementary Concept] OR iniparib OR 4-iodo-3-nitrobenzamide OR "BSI 201" OR BSI201 OR BSI-201 OR "4 iodo 3 nitrobenzamide" OR "bsi 201" OR bsi201 OR "sar 240550"	97
#8	"cediranib" [Supplementary Concept] OR cediranib OR AZD2171 OR AZD-2171 OR "AZD 2171" OR "recentin"	367
#7	"irofulven" [Supplementary Concept] OR irofulven OR 6-hydroxymethylacylfulvene OR "6-(hydroxymethyl)acylfulvene" OR "MGI 114" OR MGI.114 OR MGI-114 OR "6 hydroxymethylacylfulvene" OR HMAF OR hydroxymethylacylfulvene OR "mgi 114" OR mgi114	120
#6	"etirinotecan pegol" [Supplementary Concept] OR etirinotecan OR NKTR-102 OR "NKTR 102" OR "nktr 102"	29
#5	"enzastaurin" [Supplementary Concept] OR enzastaurin OR LY317615.HCl OR "enzastaurin hydrochloride" OR "ly 317615" OR ly317615	246
#4	"ENMD 2076" [Supplementary Concept] OR "ENMD 2076" OR ENMD-2076 OR ENMD2076	20
#3	"Rats"[Mesh] OR rat OR rats OR rattus OR "Mice"[Mesh] OR mouse OR mice OR mus OR murine OR xenograft OR xenografts OR heterograft OR heterografts OR xenogeneic OR xenogenic OR heterotransplant OR xenotransplant OR allograft OR allografts OR homograft OR homografts OR allogeneic OR allogenic OR allotransplant OR homotransplant OR alloplastic OR isograft OR isografts OR syngeneic OR "syngenic" OR isogeneic OR isogenic OR syngraft OR syngrafts OR syn-graft OR syn-grafts OR isograft OR isografts OR iso-graft OR iso-grafts OR isotransplant OR "Animals, Genetically Modified"[Mesh] OR "genetically modified" OR "genetically engineered" OR "genetically manipulated" OR "genetically-modified" OR "genetically-engineered" OR "genetically-manipulated" OR transgenic OR transgene	3450887

#2	"Drug Evaluation, Preclinical"[Mesh]	229642
#1	"Ovarian Neoplasms"[Mesh] OR ovarian OR ovary	271440

A.2 Preclinical search string used in the EMBASE database

Search	Query	N° items
#58	#1 AND #2 AND #3 AND #57	210
#57	#4 OR #5 OR #6 OR #7 OR #8 OR #9 OR #10 OR #11 OR #12 OR #13 OR #14 OR #15 OR #16 OR #17 OR #18 OR #19 OR #20 OR #21 OR #22 OR #23 OR #24 OR #25 OR #26 OR #27 OR #28 OR #29 OR #30 OR #31 OR #32 OR #33 OR #34 OR #35 OR #36 OR #37 OR #38 OR #39 OR #40 OR #41 OR #42 OR #43 OR #44 OR #45 OR #46 OR #47 OR #48 OR #49 OR #50 OR #51 OR #52 OR #53 OR #54 OR #55 OR #56	503790
#56	'paclitaxel'/exp OR paclitaxel OR 'abi 007' OR abi007 OR abraxane OR 'albumin bound paclitaxel' OR 'albumin-bound paclitaxel' OR anzatax OR apealea OR asotax OR biotax OR 'bms 181339' OR bms181339 OR 'bmy 45622' OR bmy45622 OR bristaxol OR britaxol OR coroxane OR 'dts 301' OR dts301 OR 'endotag 1' OR formoxol OR genexol OR 'genexol pm' OR hunxol OR ifaxol OR infinnium OR intaxel OR 'mbt 0206' OR mbt0206 OR medixel OR mitotax OR 'nab paclitaxel' OR 'nanoparticle albumin bound paclitaxel' OR 'nsc 125973' OR 'nsc 673089' OR nsc125973 OR nsc673089 OR 'oas pac 100' OR oaspac100 OR oncogel OR onxol OR pacitaxel OR 'paclitaxel nab' OR pacxel OR padexol OR parexel OR paxceed OR paxene OR paxus OR praxel OR 'sb 05 (terpenoid)' OR 'sb05 (terpenoid)' OR taxocris OR taxol OR 'taxus (drug)' OR taycovit OR yewtaxan) AND ('lipid nanoparticles' OR 'lipid core nanoparticles')	251
#55	'gemcitabine'/exp OR gemcitabine OR dfdcyd OR '2',2'-difluorodeoxycytidine' OR '2'-deoxy-2'-difluorocytidine' OR '2',2'-dfdc' OR '2',2'-difluoro-2'-deoxycytidine' OR '2'-deoxy-2',2'-difluorocytidine-5'-o-monophosphate' OR '2' deoxy 2', 2' difluorocytidine' OR '2', 2' difluorodeoxycytidine' OR 'd 07001' OR d07001 OR difluorodeoxycytidine OR 'ff 10832' OR ff10832 OR gemcite OR gemtro OR gemzar OR infugem OR 'ly 188011' OR ly188011	53434
#54	'lenalidomide'/exp OR lenalidomide OR '3-(4-amino-1-oxoisindolin-2-yl)piperidine-2,6-dione' OR '2,6-piperidinedione, 3-(4-amino-1,3-dihydro-1-oxo-2h- isoindol-2-yl)-' OR 'imid3 cpd' OR '3 (4 amino 1 oxo 1, 3 dihydro 2h isoindol 2 yl) 2, 6 piperidinedione' OR '3 (4 amino 1, 3 dihydro 1 oxo 2h isoindol 2 yl) glutarimide' OR '3 (4? aminoisoindoline 1? one) 1 piperidine 2, 6 dione' OR 'cc 5013' OR cc5013 OR 'cdc 501' OR 'cdc 5013' OR cdc501 OR cdc5013 OR 'enmd 0997' OR enmd0997 OR 'imid 3' OR imid3 OR revimid OR revlimid	17576
#53	'birinapant'/exp OR birinapant OR 'n, n? [(6, 6? difluoro 1h, 1?h 2, 2? biindolyl 3, 3? diyl) bis [methylene (4 hydroxypyrrolidine 2, 1 diyl) (1 ethyl 2 oxoethylene)]] bis [2 (methylamino) propanamide]' OR 'n, n? [(6, 6? difluoro [1h, 1?h 2, 2? biindole] 3, 3? diyl) bis [methylene (4 hydroxypyrrolidine 2, 1 diyl) (1 oxobutane 1, 2 diyl)]] bis [2 (methylamino) propanamide]' OR 'n, n? [(6, 6? difluoro [2, 2? bi 1h indole] 3, 3? diyl) bis [methylene (4 hydroxy 2, 1 pyrrolidinediyl) (1 ethyl 2 oxo 2, 1 ethanedyl)]] bis [2 (methylamino) propanamide]' OR 'tl 32711' OR tl32711	222
#52	'belinostat'/exp OR belinostat OR '3 phenylsulfamoylcinnamohydroxamic acid' OR beleodaq OR 'n hydroxy 3 [(3 phenylsulfamoyl) phenyl] 2 propenamide' OR 'n hydroxy 3 [3 (n phenylsulfamoyl) phenyl] prop 2 enamide' OR 'pdx 101' OR pdx101 OR 'pxd 101' OR pxd101 OR belecodaq	1373
#51	'nintedanib'/exp OR nintedanib OR '2, 3 dihydro 3 [[[4 [methyl 2 (4 methyl 1 piperazinyl) acetyl] amino] phenyl] amino] phenylmethylene] 2 oxo 1h indole 6 carboxylic acid methyl ester' OR '2, 3 dihydro 3 [[[4 [n methyl 2 (4 methyl 1 piperazinyl) acetamido] phenyl] amino] (phenyl) methylidene] 2 oxo 1h indole 6 carboxylic acid methyl ester' OR intedanib OR 'methyl 3 [[[4 [n methyl 2 (4 methylpiperazin 1 yl) acetamido] phenyl] amino] (phenyl) methylidene] 2 oxo 2, 3 dihydro 1h indole 6 carboxylate' OR 'nintedanib esylate' OR ofev OR vargatef OR bibf1120 OR 'bibf 1120'	2610

#50	'selumetinib'/exp OR selumetinib OR '5 (4 bromo 2 chloroanilino) 4 fluoro 1 methyl 1h benzimidazole 6 carbohydroxamic acid 2 hydroxyethyl ester' OR '5 (4 bromo 2 chlorophenylamino) 4 fluoro 1 methyl 1h benzimidazole 6 carbohydroxamic acid 2 hydroxyethyl ester' OR '5 [(4 bromo 2 chlorophenyl) amino] 4 fluoro n (2 hydroxyethoxy) 1 methyl 1h benzimidazole 6 carboxamide' OR 'arry 142886' OR arry142886 OR 'azd 6244' OR azd6244 OR 'selumetinib sulfate' OR 'selumetinib sulphate'	2791
#49	'aflibercept'/exp OR aflibercept OR 'vegfr trap - regeneron' OR 'vegfr trap-eye' OR eylea OR zaltrap OR ave0005 OR 'ave 0005' OR ave005 OR 'ave 005' OR 'vascular endothelial growth factor trap' OR 'vasculotropin trap' OR 'vegfr trap' OR 'ziv aflibercept'	5301
#48	'vandetanib'/exp OR vandetanib OR 'n-(4-bromo-2-fluorophenyl)-6-methoxy-7-((1-methylpiperidin-4-yl)methoxy)quinazolin-4-amine' OR 'azd 6474' OR azd6474 OR caprelsa OR 'n (4 bromo 2 fluorophenyl) 6 methoxy 7 (1 methyl 4 piperidinylmethoxy) 4 quinazolinamine' OR 'n (4 bromo 2 fluorophenyl) 6 methoxy 7 (1 methylpiperidin 4 ylmethoxy) quinazolin 4 amine' OR vandetanib OR zactima OR 'zd 6474' OR zd6474	4442
#47	'prexasertib'/exp OR prexasertib OR '5 [[5 [2 (3 aminopropoxy) 6 methoxyphenyl] 1h pyrazol 3 yl] amino] 2 pyrazinecarbonitrile' OR '5 [[5 [2 (3 aminopropoxy) 6 methoxyphenyl] 1h pyrazol 3 yl] amino] pyrazine 2 carbonitrile' OR 'ly 2606368' OR ly2606368	147
#46	'trabectedin'/exp OR trabectedin OR 'nsc 684766' OR 'ecteinascidin 743' OR 'et 743' OR et743 OR yondelis	2425
#45	'gimatecan'/exp OR '7-t-butoxyiminomethylcamptothecin' OR '7-t-butoxyiminomethylcamptothecin' OR gimatecan OR '7 tert butoxyiminomethylcamptothecin' OR 'cpt 184' OR cpt184 OR 'lbq 707' OR lbq707 OR 'st 1481' OR st1481	144
#44	'dasatinib'/exp OR dasatinib OR 'n-(2-chloro-6-methylphenyl)-2-(6-(4-(2-hydroxyethyl)piperazin-1-yl)-2-methylpyrimidin-4-ylamino)thiazole-5-carboxamide' OR '(18f)-n-(2-chloro-6-methylphenyl)-2-(6-(4-(2-hydroxyethyl)piperazin-1-yl)-2-methylpyrimidin-4-ylamino)thiazole-5-carboxamide' OR '354825, bms' OR 'bms 354825' OR 'bms 354825 03' OR 'bms 354825-03' OR 'bms 35482503' OR bms354825 OR 'bms354825 03' OR bms35482503 OR 'dasatinib hydrate' OR 'n (2 chloro 6 methylphenyl) 2 [6 [4 (2 hydroxyethyl) 1 piperazinyl] 2 methyl 4 pyrimidinylamino] 5 thiazolecarboxamide' OR 'n (2 chloro 6 methylphenyl) 2 [6 [4 (2 hydroxyethyl) piperazin 1 yl] 2 methylpyrimidin 4 ylamino] thiazole 5 carboxamide' OR 'n (2 chloro 6 methylphenyl) 2 [[6 [4 (2 hydroxyethyl) 1 piperazinyl] 2 methyl 4 pyrimidinyl] amino] 5 thiazolecarboxamide' OR 'n (2 chloro 6 methylphenyl) 2 [[6 [4 (2 hydroxyethyl) piperazin 1 yl] 2 methylpyrimidin 4 yl] amino] thiazole 5 carboxamide' OR sprycel	12579
#43	'ixabepilone'/exp OR ixabepilone OR '7, 11 dihydroxy 8, 8, 10, 12, 16 pentamethyl 3 [1 methyl 2 (2 methyl 4 thiazolyl) ethenyl] 17 oxa 4 azabicyclo [14.1.0] heptadecane 5, 9 dione' OR 'azaepothilone b' OR 'bms 247550' OR 'bms 247550 1' OR 'bms 247550-1' OR bms247550 OR 'bms247550 1' OR ixempra OR 'ixempra kit' OR 'nsc 710428' OR nsc710428	1749
#42	'perifosine'/exp OR perifosine OR 'octadecyl-(1,1-dimethyl-4-piperidyl)phosphate' OR '4 [[hydroxy (octadecyloxy) phosphinyl] oxy] 1, 1 dimethylpiperidinium' OR 'd 21266' OR d21266 OR 'krx 0401' OR krx0401 OR 'nka 17' OR nka17 OR 'octadecyl (1, 1 dimethylpiperidinio 4 yl) phosphate'	1274
#41	'rivoceranib'/exp OR rivoceranib OR aitan OR apatinib OR 'apatinib mesilate' OR 'apatinib mesylate' OR 'apatinib methanesulfonate' OR 'n [4 (1 cyanocyclopentyl) phenyl] 2 (4 pyridinylmethyl) amino 3 pyridinecarboxamide' OR 'n [4 (1 cyanocyclopentyl) phenyl] 2 [[(4 pyridinyl) methyl] amino] 3 pyridinecarboxamide' OR 'n [4 (1 cyanocyclopentyl) phenyl] 2 [[(pyridin 4 yl) methyl] amino] pyridine 3 carboxamide' OR 'rivoceranib mesilate' OR 'rivoceranib mesylate' OR 'rivoceranib methanesulfonate' OR 'yn 968d1' OR yn968d1	684
#40	'volasertib'/exp OR volasertib OR 'bi 6727' OR bi6727 OR 'n [4 [4 (cyclopropylmethyl) 1 piperazinyl] cyclohexyl] 4 [(7 ethyl 5, 6, 7, 8 tetrahydro 5 methyl 8 (1 methylethyl) 6 oxo 2 pteridinyl) amino] 3 methoxybenzamide' OR 'n [4 [4 (cyclopropylmethyl) 1 piperazinyl] cyclohexyl] 4 [(7 ethyl 5, 6, 7, 8 tetrahydro 8 isopropyl 5 methyl 6 oxo 2 pteridinyl) amino] 3 methoxybenzamide' OR 'n [4 [4 (cyclopropylmethyl) piperazin 1 yl] cyclohexyl] 4 [[7 ethyl 5	421

	methyl 6 oxo 8 (propan 2 yl) 5, 6, 7, 8 tetrahydropteridin 2 yl] amino] 3 methoxybenzamide' OR 'n [4 [4 (cyclopropylmethyl) piperazin 1 yl] cyclohexyl] 4 [[7 ethyl 5 methyl 8 (1 methylethyl) 6 oxo 5, 6, 7, 8 tetrahydropteridin 2 yl] amino] 3 methoxybenzamide' OR 'volasertib hydrochloride' OR 'volasertib trihydrochloride'	
#39	'temsirolimus'/exp OR temsirolimus OR 'rapamycin, 42-(3-hydroxy-2-(hydroxymethyl)-2-methylpropanoate)' OR '42 o [2, 2 bis (hydroxymethyl) propionyl] rapamycin' OR 'cci 779' OR cci779 OR 'cell cycle inhibitor 779' OR 'nsc 683864' OR nsc683864 OR 'rapamycin 2, 2 bis (hydroxymethyl) propionate' OR 'rapamycin 42 [2, 2 bis (hydroxymethyl) propionate]' OR torisel OR 'way-cci 779'	7877
#38	'imatinib'/exp OR 'mesylate, imatinib' OR 'imatinib methanesulfonate' OR 'methanesulfonate, imatinib' OR imatinib OR 'alpha-(4-methyl-1-piperazinyl)-3?-(4-(3-pyridyl)-2-pyrimidinyl)amino)-p-tolu-p-toluidide' OR '2 [2 methyl 5 [4 (4 methyl 1 piperazinylmethyl) benzamido] anilino] 4 (3 pyridyl) pyrimidine' OR '4 (4 methylpiperazin 1 ylmethyl) n [4 methyl 3 [4 (3 pyridyl) pyrimidin 2 ylamino] phenyl] benzamide' OR '4 [(4 methyl 1 piperazinyl) methyl] n [4 methyl 3 [[4 (3 pyridinyl) 2 pyrimidinyl] amino] phenyl] benzamide' OR 'alpha (4 methyl 1 piperazinyl) 3? [[4 (3 pyridyl) 2 pyrimidinyl] amino] para tolu para toluidide' OR 'cgp 57148' OR 'cgp 57148b' OR cgp57148 OR cgp57148b OR gleevac OR gleevec OR glivec OR glivic OR 'imatinib mesilate' OR 'imatinib mesylate' OR ruvise OR 'signal transduction inhibitor 571' OR 'st 1571' OR st1571 OR 'sti 571' OR sti571	41122
#37	'alisertib'/exp OR alisertib OR '4 [[9 chloro 7 (2 fluoro 6 methoxyphenyl) 5h pyrimido [5, 4 d] [2] benzazepin 2 yl] amino] 2 methoxybenzoic acid' OR 'alisertib sodium' OR 'mln 8237' OR 'mln 8237 004' OR 'mln 8237-004' OR mln8237 OR 'mln8237 004'	985
#36	'cisplatin'/exp OR 'liposomal cisplatin' OR 'spi 077' OR spi077 OR 'spi 77' OR 'stealth liposomal cisplatin' OR 'cisplatin liposomal' OR (cisplatin AND (liposomes OR liposomal))	171244
#35	'lurbinectedin'/exp OR lurbinectedin OR '8, 14 dihydroxy 6?, 9 dimethoxy 4, 10, 23 trimethyl 19 oxo 2?, 3?, 4?, 6, 7, 9?, 12, 13, 14, 16 decahydro 6ah spiro [7, 13 azano 6, 16 (epithiopropanooxymethano) [1, 3] dioxolo [7, 8] isoquinolino [3, 2 b] [3] benzazocine 20, 1? pyrido [3, 4 b] indol] 5 yl acetate' OR 'pm 01183' OR 'pm 1183' OR pm01183 OR pm1183	151
#34	'dalantercept'/exp OR 'alk1-fc fusion protein, human' OR dalantercept OR 'ace 041' OR ace041	79
#33	'olaparib'/exp OR olaparib OR azd221 OR '1 (cyclopropylcarbonyl) 4 [2 fluoro 5 [(4 oxo 3, 4 dihydrophthalazin 1 yl) methyl] benzoyl] piperazine' OR '4 [3 (4 cyclopropanecarbonylpiperazine 1 carbonyl) 4 fluorobenzyl] 2h phthalazin 1 one' OR '4 [[3 [[4 (cyclopropylcarbonyl) 1 piperazinyl] carbonyl] 4 fluorophenyl] methyl] 1 (2h) phthalazinone' OR '4 [[3 [[4 (cyclopropylcarbonyl) piperazin 1 yl] carbonyl] 4 fluorophenyl] methyl] phthalazin 1 (2h) one' OR 'azd 2281' OR azd2281 OR 'ku 0059436' OR 'ku 59436' OR ku0059436 OR ku59436 OR lynparza	3895
#32	'doxorubicin'/exp OR doxorubicin OR ribodoxo OR 'doxo cell' OR 'urokit doxo-cell' OR 'urokit doxo cell' OR 'doxorubicina ferrer farm' OR 'doxorubicina funk' OR 'doxorubicina tedec' OR 'doxorubicine baxter' OR doxotec OR onkodox OR '14 hydroxydaunomycin' OR '14 hydroxydaunorubicin' OR a.d.mycin OR adriablastin OR adriablastina OR 'adriablastina r.d.' OR adriablastine OR adriacin OR adriamicina OR adriamicine OR adriamycin OR 'adriamycin hydrochloride' OR 'adriamycin p.f.s.' OR 'adriamycin pfs' OR 'adriamycin r.d.f.' OR 'adriamycin rd' OR 'adriamycin rdf' OR adriamycina OR adriblastin OR adriblastina OR 'adriblastina cs' OR 'adriblastina pfs' OR adriblastine OR adrim OR adrimedac OR adrubicin OR amminac OR caelix OR caelyx OR 'caelyx/doxil' OR carcinocin OR dexorubicin OR 'dox sl' OR doxil OR 'doxil (liposomal)' OR doxolem OR 'doxor lyo' OR 'doxorubicin hydrochloride' OR 'doxorubicin meiji' OR 'doxorubicin, liposomal' OR doxorubicine OR doxorubin OR evacet OR farmiblastina OR 'fi 106' OR fi106 OR ifadox OR lipodox OR 'liposomal doxorubicin' OR 'mcc 465' OR mcc465 OR myocet OR 'nsc 123127' OR nsc123127 OR 'pegylated liposomal doxorubicin' OR 'polyethylene glycol-coated liposomal doxorubicin' OR rastocin OR resmycin OR 'rp 25253' OR rp25253 OR rubex OR rubidox OR sarcodoxome OR 'tlc d 99'	186007

#31	'4 (8 cyclopentyl 7 ethyl 5,6,7,8 tetrahydro 5 methyl 6 oxo 2 pteridinylamino) 3 methoxy n (1 methyl 4 piperidinyl)benzamide'/exp OR '4 (8 cyclopentyl 7 ethyl 5,6,7,8 tetrahydro 5 methyl 6 oxo 2 pteridinylamino) 3 methoxy n (1 methyl 4 piperidinyl)benzamide' OR bi2536 OR 'bi 2536'	512
#30	'bendamustine'/exp OR 'hydrochloride, bendamustine' OR bendamustine OR bendamustin OR '4 [5 [bis (2 chloroethyl) amino] 1 methylbenzimidazol 2 yl] butyric acid' OR '5 [bis (2 chloroethyl) amino] 1 methyl 2 benzimidazolebutyric acid' OR 'bendamustine hydrochloride' OR bendeka OR 'cimet 3393' OR cytotasan OR 'cytotasan r' OR cytotasane OR 'imet 3393' OR levact OR ribomustin OR ribovact OR treanda OR 'zimet 3393' OR zimet3393	5940
#29	'eribulin'/exp OR eribulin OR nsc707389 OR 'nsc 707389' OR 'b 1793' OR 'b-1793' OR 'b 1939' OR 'eribulin monomethanesulfonate' OR 'eribulin (as mesylate)' OR 'e 7389' OR e7389 OR 'er 086526' OR 'er 86526' OR er086526 OR er86526 OR 'eribulin mesilate' OR 'eribulin mesylate' OR halaven	2098
#28	'danusertib'/exp OR danusertib OR '4 (4 methyl 1 piperazinyl) n [1, 4, 5, 6 tetrahydro 5 (2 methoxy 2 phenylacetyl) pyrrolo [3, 4 c] pyrazol 3 yl] benzamide' OR 'n [5 (2 methoxy 2 phenylacetyl) 1, 4, 5, 6 tetrahydropyrrolo [3, 4 c] pyrazol 3 yl] 4 (4 methylpiperazin 1 yl) benzamide' OR 'pha 739358' OR pha739358	355
#27	'zoptarelin doxorubicin'/exp OR 'lhrh, lysine(6)-doxorubicin' OR 'luteinizing hormone-releasing factor (pig), 6-(n6-(5-(2-(4- ((3-amino-2,3,6-trideoxy-alpha-l-lyxo-hexopyranosyl)oxy)-1,2,3,4,6,11-hexahydro-2,5,12-trihydroxy-7-methoxy-6,11-dioxo-2-naphthacenyl)-2-oxoethoxy)-1,5-dioxopentyl)-d- lysine)-, (2s-cis)-' OR 'zoptarelin doxorubicin' OR 'lys(6)-lhrh-doxorubicin' OR 'aezs 108' OR aezs108 OR 'an 152' OR an152 OR 'd 81858' OR d81858 OR 'doxorubicin lhrh [6 dextro lysine]' OR 'gonadorelin [6 dextro lysine] doxorubicin' OR 'lhrh [6 dextro lysine] doxorubicin' OR 'zen 008' OR zen008 OR 'zopatrelin doxorubicin acetate'	155
#26	'sorafenib'/exp OR sorafenib OR nexavar OR 'sorafenib n-oxide' OR 'sorafenib n oxide' OR 'bay 673472' OR 'bay 545-9085' OR 'bay 5459085' OR 'bay 545 9085' OR bay5459085 OR 'sorafenib tosylate' OR '4-(4-(3-(4-chloro-3-trifluoromethylphenyl)ureido)phenoxy)pyridine-2-carboxylic acid methanamide-4-methylbenzenesulfonate' OR '4 [4 [3 [4 chloro 3 (trifluoromethyl) phenyl] ureido] phenoxy] n methyl 2 pyridinecarboxamide' OR 'bay 43 9006' OR 'bay 43-9006' OR 'bay 439006' OR bay43) AND 9006 OR 'bay43 9006' OR bay439006 OR nexavar OR 'sorafenib tosylate'	3522
#25	'cabozantinib'/exp OR cabozantinib OR 'xl184 cpd' OR 'bms 907351' OR bms907351 OR cabometyx OR 'cabozantinib malate' OR 'cabozantinib s malate' OR 'cabozantinib s-malate' OR cometriq OR 'cyclopropane 1, 1 dicarboxylic acid [4 (6, 7 dimethoxyquinolin 4 yloxy) phenyl] amide (4 fluorophenyl) amide' OR 'n [4 (6, 7 dimethoxy 4 quinolinyl oxy) phenyl] n? (4 fluorophenyl) 1, 1 cyclopropanedicarboxamide' OR 'n [4 [(6, 7 dimethoxyquinolin 4 yl) oxy] phenyl] n? (4 fluorophenyl) cyclopropane 1, 1 dicarboxamide' OR 'xl 184' OR xl184	3049
#24	'pazopanib'/exp OR pazopanib OR gw780604 OR 'gw 780604' OR '5 [[4 [(2, 3 dimethyl 2h indazol 6 yl) methylamino] 2 pyrimidinyl] amino] 2 methylbenzenesulfonamide' OR 'armala' OR 'gw 786034' OR 'gw 786034b' OR 'gw 786034x' OR gw786034 OR gw786034b OR gw786034x OR 'pazopanib hydrochloride' OR 'sb 710468' OR 'sb 710468a' OR sb710468 OR sb710468a OR votrient	7043
#23	'sagopilone'/exp OR sagopilone OR 'de 03757' OR 'epo 477' OR 'sh y 03757' OR 'bay 86 5302' OR 'sh y03757' OR 'sh y03757a' OR shy03757 OR shy03757a OR 'zk 219477' OR 'zk epo' OR 'zk epothilone' OR zk219477	187
#22	'topotecan'/exp OR topotecan OR '9-dimethylaminomethyl-10-hydroxycamptothecin' OR 'hydrochloride, topotecan' OR 'nogitecan hydrochloride' OR 'hydrochloride, nogitecan' OR 'topotecan monohydrochloride, (s)-isomer' OR 'sk and f-104864-a' OR 'sk and f 104864 a' OR 'sk and f104864a' OR hycamtamine OR '9 dimethylaminomethyl 10 hydroxycamptothecin' OR 'e 89001' OR e89001 OR evotopin OR hycamtamine OR hycamtin OR lutecan OR 'nsc 609699' OR nsc609699 OR oncotecan OR potactasol OR ribocamtin OR 'skf 104864' OR 'skf 104864 a' OR 'skf 104864a' OR 'skf s 104864a' OR 'skf s104864a' OR skf104864 OR skf104864a OR topocan OR topoliquid OR 'topotecan hydrochloride' OR topotecane OR topotecano OR topotekan OR topotel OR topovin	11475

#21	'n (6,7 dihydro 6 oxo 5h dibenz[b,d]azepin 7 yl) 2,2 dimethyl n` (2,2,3,3,3 pentafluoropropyl)propanediamide'/exp OR 'n (6,7 dihydro 6 oxo 5h dibenz[b,d]azepin 7 yl) 2,2 dimethyl n` (2,2,3,3,3 pentafluoropropyl)propanediamide' OR ro4929097 OR 'ro 4929097'	280
#20	'epothilone b'/exp OR 'epothilone b' OR patupilone OR 'epothilon b' OR 'epo 906' OR epo906	1196
#19	peptide AND 'a6' OR ('urokinase plasminogen activator' AND '136-143') OR 'urokinase derived peptide' OR 'urokinase-derived peptide' OR 'ac lys pro ser ser pro pro glu glu nh2' OR 'ac kpssppee nh2' OR 'acetyl-lysyl-prolyl-seryl-seryl-prolyl-prolyl-glutamyl-glutamic acid amide'	378
#18	'tasquinimod'/exp OR tasquinimod OR '4-hydroxy-5-methoxy-n,1-dimethyl-2-oxo-n-((4-trifluoromethyl)phenyl)-1,2-dihydroquinoline-3-carboxamide' OR '4 hydroxy 5 methoxy n, 1 dimethyl 2 oxo n [4 (trifluoromethyl) phenyl] 1, 2 dihydroquinoline 3 carboxamide' OR 'abr 215050' OR abr215050	221
#17	'veliparib'/exp OR veliparib OR '2-((r)-2-methylpyrrolidin-2-yl)-1h-benzimidazole-4-carboxamide' OR '2-(2-methylpyrrolidin-2-yl)-1h-benzimidazole-4-carboxamide' OR '2 (2 methyl 2 pyrrolidinyl) 1h benzimidazole 4 carboxamide' OR '2 (2 methylpyrrolidin 2 yl) 1h benzimidazole 4 carboxamide' OR 'abt 888' OR abt888	1828
#16	'sunitinib'/exp OR sunitinib OR '5-(5-fluoro-2-oxo-1,2-dihydroindolidenemethyl)-2,4-dimethyl-1h-pyrrole-3-carboxylic acid (2-diethylaminoethyl)amide' OR '5 (5 fluoro 1, 2 dihydro 2 oxo 3 indolylidenemethyl) 2, 4 dimethyl 1h pyrrole 3 carboxylic acid (2 diethylaminoethyl) amide' OR '5 (5 fluoro 2 oxo 1, 2 dihydroindol 3 ylidenemethyl) 2, 4 dimethyl 1h pyrrole 3 carboxylic acid (2 diethylaminoethyl) amide' OR 'n [2 (diethylamino) ethyl] 5 [(5 fluoro 1, 2 dihydro 2 oxo 3h indol 3 ylidene) methyl] 2, 4 dimethyl 1h pyrrole 3 carboxamide' OR 'pha 2909040ad' OR pha2909040ad OR 'su 010398' OR 'su 011248' OR 'su 10398' OR 'su 11248' OR su010398 OR su011248 OR su10398 OR su11248 OR 'sunitinib malate' OR 'suo 11248' OR suo11248 OR sutent	21503
#15	'paclitaxel'/exp OR '130-nm albumin-bound paclitaxel' OR (albumin AND paclitaxel) OR 'paclitaxel, albumin-bound' OR 'protein-bound paclitaxel' OR 'paclitaxel, protein-bound' OR 'protein bound paclitaxel' OR 'abi 007' OR abi007 OR abraxane OR 'albumin bound paclitaxel' OR 'albumin-bound paclitaxel' OR anzatax OR apealea OR asotax OR biotax OR 'bms 181339' OR bms181339 OR 'bmy 45622' OR bmy45622 OR bristaxol OR britaxol OR coroxane OR 'dts 301' OR dts301 OR 'endotag 1' OR formoxol OR genexol OR 'genexol pm' OR hunxol OR ifaxol OR infinnium OR intaxel OR 'mbt 0206' OR mbt0206 OR medixel OR mitotax OR 'nab paclitaxel' OR 'nanoparticle albumin bound paclitaxel' OR 'nsc 125973' OR 'nsc 673089' OR nsc125973 OR nsc673089 OR 'oas pac 100' OR oaspac100 OR oncogel OR onxol OR pacitaxel OR 'paclitaxel nab' OR pacxel OR padexol OR parexel OR paxceed OR paxene OR paxus OR praxel OR 'sb 05 (terpenoid)' OR 'sb05 (terpenoid)' OR taxocris OR taxol OR 'taxus (drug)' OR taycovit OR yewtaxan	101187
#14	'motesanib'/exp OR motesanib OR 'amg 706' OR amg706 OR 'motesanib diphosphate' OR 'n (2, 3 dihydro 3, 3 dimethyl 6 indolyl) 2 (4 pyridinylmethylamino) 3 pyridinecarboxamide' OR 'n (3, 3 dimethyl 2, 3 dihydro 1h indol 6 yl) 2 (pyridin 4 ylmethylamino) pyridine 3 carboxamide' OR 'n (3, 3 dimethyl 6 indolyl) 2 (4 pyridinylmethylamino) nicotinamide'	967
#13	'lapatinib'/exp OR lapatinib OR 'n-(3-chloro-4-(((3-fluorobenzyl)oxy)phenyl)-6-(5-(((2-methylsulfonyl)ethyl)amino)methyl) -2-furyl)-4-quinazolinamine' OR gw282974x OR 'gw 282974x' OR '4 [3 chloro 4 (3 fluorobenzyloxy) anilino] 6 [5 [2 (methylsulfonyl) ethylaminomethyl] 2 furyl] quinazoline' OR 'gw 2016' OR 'gw 572016' OR 'gw 572016f' OR gw2016 OR gw572016 OR gw572016f OR 'lapatinib ditosylate' OR 'lapatinib ditosylate monohydrate' OR 'lapatinib tosylate' OR 'n [3 chloro 4 [(3 fluorobenzyl) oxy] phenyl] 6 [5 [[2 (methylsulfonyl) ethyl] amino] methyl] furan 2 yl] quinazolin 4 amine bis (4 methylbenzenesulfonate)' OR tykerb OR tyverb	11411
#12	'asparaginase'/exp OR asparaginase OR 'monomethoxypolyethylene glycol-conjugated asparaginase' OR 'peg(2)-asp' OR '2,4-bis(2-methoxypolyethyleneglycol)-6-chloro-s-triazine-conjugated l-asparaginase' OR 'asparaginase 2' OR 'asparaginase a' OR 'asparaginase ag' OR 'asparaginase b' OR 'asparaginase erwinia chrysanthemi' OR 'asparagine amidohydrolase' OR aspariginase OR aspariginase OR colaspase OR collaspase OR crasnitin OR crisantaspase OR 'e.c.	15632

	3.5.1.1' OR elaspar OR elspar OR erwinase OR erwinaze OR 'ery 001' OR 'ery asp' OR ery001 OR eryasp OR eryaspase OR 'fb b 6366' OR 'fb b6366' OR fbb6366 OR graspa OR kidrolase OR krasnitin OR 'l asparaginase' OR 'l asparaginase a' OR 'l asparagine amidohydrolase' OR 'l asparaginase' OR laspar OR 'levo asparaginase' OR 'levo asparagine amidohydrolase' OR 'nsc 109229' OR nsc109229 OR paronal OR 'asparagine deaminase' OR 'deaminase, asparagine' OR 'asparaginase ii' OR leunase OR 'asparaginase medac' OR 'medac, asparaginase'	
#11	'rucaparib'/exp OR rucaparib OR '8 fluoro 1, 3, 4, 5 tetrahydro 2 [4 (methylaminomethyl) phenyl] 6h pyrrolo [4, 3, 2 ef] [2] benzazepin 6 one' OR '8 fluoro 1, 3, 4, 5 tetrahydro 2 [4 [(methylamino) methyl] phenyl] 6h azepino [5, 4, 3 cd] indol 6 one' OR '8 fluoro 2 [4 [(methylamino) methyl] phenyl] 1, 3, 4, 5 tetrahydro 6h azepino [5, 4, 3 cd] indol 6 one' OR '8 fluoro 2 [4 [(methylamino) methyl] phenyl] 1, 3, 4, 5 tetrahydro 6h pyrrolo [4, 3, 2 ef] [2] benzazepin 6 one' OR '8 fluoro 3, 4 dihydro 2 [4 (methylaminomethyl) phenyl] pyrrolo [3, 4, 5 e, f] [2] benzazepin 6 (5h) one' OR '8 fluoro 3, 4 dihydro 2 [4 (methylaminomethyl) phenyl] pyrrolo [4, 3, 2 ef] [2] benzazepin 6 (5h) one' OR 'ag 014699' OR 'ag 14447' OR 'ag 14699' OR ag014699 OR ag14447 OR ag14699 OR 'co 338' OR co338 OR 'pf 01367338' OR 'pf 1367338' OR 'pf 1367338 bw' OR pf01367338 OR pf1367338 OR pf1367338bw OR rubraca OR 'rucaparib camphorsulfonate' OR 'rucaparib camsylate' OR 'rucaparib camsylate' OR 'rucaparib phosphate'	1035
#10	'elacytarabine'/exp OR '5?-oleoyl cytarabine' OR '5?-oleyl-ara-c' OR '5?-oleoyl cytosine arabinoside' OR elacytarabine OR '4 amino 1 [5 o (octadec 9 enoyl) beta dextro arabinofuranosyl] pyrimidin 2 (1h) one' OR '5? o (9?? octadecenoyl) 1 beta dextro arabinofuranosylcytosine' OR '5? o (trans 9?? octadecenoyl) 1 beta d arabinofuranosylcytosine' OR 'cp 4055' OR cp4055 OR 'cytarabine 5? elaidic acid ester' OR elacyt	106
#9	'iniparib'/exp OR iniparib OR '4 iodo 3 nitrobenzamide' OR 'bsi 201' OR bsi201 OR 'sar 240550' OR sar240550	749
#8	'cediranib'/exp OR cediranib OR '4-((4-fluoro-2-methyl-1h-indol-5-yl)oxy)-6-methoxy-7-(3-(pyrrolidin-1-yl)propoxy)quinazoline' OR '4 (4 fluoro 2 methyl 5 indolyloxy) 6 methoxy 7 [3 (1 pyrrolidinyl) propoxy] quinazoline' OR '4 [(4 fluoro 2 methyl 1h indol 5 yl) oxy] 6 methoxy 7 [3 (pyrrolidin 1 yl) propoxy] quinazoline' OR 'azd 2171' OR azd2171 OR 'cediranib maleate' OR recentin OR zemfirza	2688
#7	'irofulven'/exp OR irofulven OR 'hmaf cpd' OR '6-(hydroxymethyl)acylfulvene' OR mgi.114 OR '6 hydroxymethylacylfulvene' OR '6? hydroxy 3? hydroxymethyl 2?, 4?, 6? trimethylspiro [cyclopropane 1, 5? 5h inden] 7? (6?h) one' OR hmaf OR hydroxymethylacylfulvene OR irofulvene OR 'mgi 114' OR mgi114 OR 'nsc 683863' OR nsc683863	226
#6	'etirinotecan pegol'/exp OR etirinotecan OR 'etirinotecan pegol tetrahydrochloride' OR 'etirinotecan pegol tetratriflutate' OR 'nktr 102' OR nktr102 OR onzeald	138
#5	'enzastaurin'/exp OR enzastaurin OR ly317615.hcl OR '3 (1 methyl 1h indol 3 yl) 4 [1 [1 (pyridin 2 ylmethyl) piperidin 4 yl] 1h indol 3 yl] 1h pyrrole 2, 5 dione' OR '3 (1 methyl 1h indol 3 yl) 4 [1 [1 (pyridin 2 ylmethyl) piperidin 4 yl] 1h indol 3 yl] pyrrole 2, 5 dione' OR '3 (1 methyl 3 indolyl) 4 [1 [1 (2 pyridinylmethyl) 4 piperidinyl] 3 indolyl] 2, 5 pyrroledione' OR 'enzastaurin hydrochloride' OR 'ly 317615' OR ly317615	1035
#4	'enmd 2076'/exp OR 'enmd 2076' OR enmd2076	115
#3	'rat'/exp OR rat OR rats OR rattus OR 'mouse'/exp OR mouse OR mice OR mus OR murine OR xenograft OR xenografts OR heterograft OR heterografts OR xenogeneic OR xenogenic OR heterotransplant OR xenotransplant OR allograft OR allografts OR homograft OR homografts OR allogeneic OR allogenic OR allotransplant OR homotransplant OR alloplastic OR syngeneic OR syngenic OR isogeneic OR isogenic OR syngraft OR syngrafts OR 'syn graft' OR 'syn grafts' OR isograft OR isografts OR 'iso graft' OR 'iso grafts' OR 'iso transplant' OR isotransplant OR 'genetically engineered mouse strain'/exp OR 'genetically engineered rat strain'/exp OR 'genetically modified' OR 'genetically engineered' OR 'genetically manipulated' OR 'genetically-modified' OR 'genetically-engineered' OR 'genetically-manipulated' OR transgenic OR transgene	4110903
#2	'drug screening'/exp	172453

#1	'ovary tumor'/exp OR ovarian OR ovary	365747
----	---------------------------------------	--------

Appendix B

➤ Repetition and external validity

- α1.** Is the same experiment repeated more than once by a single lab? Yes No
 N° of repetition _____
 Reasons for repetition _____
- α2.** N° of different species in which the EOC experiment has been repeated N° _____
- α3.** N° of different cancer models in which the EOC experiment has been repeated N° _____
- α4.** N° of participating laboratories Monolab Multilab

➤ Internal validity

- β1.** Is an internal control group used? Yes No
- β1 a.** Are animals randomly allocated to treatments? Yes No Unknown
 Alternative method to randomization _____
- β1 b.** If randomization is used, is the randomization method stated? Yes No
 Randomization method _____
- β1 c.** If randomization is used, is allocation concealment employed? Yes No Unknown
- β1 d.** Are the animals randomly housed within the animal room? Yes No Unknown
- β1 e.** Are there equal numbers per treatment group? Yes No Unknown
- β1 f.** If not, is this justified? Yes No Not applicable
 Specify justification _____
- β2.** Is the experimental unit clearly identified? Yes No
 Specify experimental unit _____
 Suspicious cause of pseudo-replication _____
- β3.** Role of CC monotherapy arm Active Control Both

➤ Statistical design

- γ1.** Is an active control group used? Yes No
- γ2.** Number of design factors N° _____
- γ3.** Number of treatment (design) groups N° _____
- γ4.** Is a factorial design used? Yes No
- γ5.** Is dose-response (i.e. ≥ 3 doses) evaluated? Yes No
- γ6.** Is blocking used? Yes No Unknown
- γ7.** Type of experiment Between units Within units Both

➤ **Sample size**

- δ1.** Is the sample size justified? Yes No
- δ1 a.** If yes, specify method
- | | | |
|-------------------|-------|----|
| common sense | Yes | No |
| power analysis | Yes | No |
| resource equation | Yes | No |
| other | _____ | |
- δ2.** Total number of enrolled animals per arm (methods section) Exact Estimate Unknown
- specify N° _____

➤ **Outcomes and their assessment**

- ε1.** Can the primary outcomes of antitumor activity clearly be identified? Yes No

- ε2.** N° Name Definition

Primary		
1		
2		
3		
4		
5		

- ε3.** Is the antitumor activity endpoint (i.e. event) clearly defined? Yes No

Specify definition _____

- ε4.** Are the competing events (compEv) clearly defined? Yes No

Specify compEv n.1 _____

compEv n.2 _____

compEv n.3 _____

- ε5.** Outcome assessment

- ε5 a.** Are the caregivers and/or investigators blinded from knowledge of which intervention each animal received during the experiment? Yes No Unknown

- ε5 b.** Are animals selected at random for outcome assessment? Yes No Unknown

- ε5 c.** Is the outcome assessor blinded? Yes No Unknown

➤ **Statistical analysis**

- ζ1.** Are inferential methods used to demonstrate antitumor activity? Yes No

- ζ1 a.** Hypothesis test Yes No

method n°1 _____

method n°2 _____

method n°3 _____

ζ1 b. Estimation Yes No

method n°1 _____

method n°2 _____

method n°3 _____

ζ2. Are descriptive methods used to demonstrate antitumor activity? Yes No

method n°1 _____

method n°2 _____

method n°3 _____

ζ3. Are methods for correction of multiple comparison used? Yes No Not applicable

method n°1 _____

method n°2 _____

method n°3 _____

ζ4. Are statistical assumptions used to analyze tumor growth data justified?

Yes No Not applicable

method n°1 _____

method n°2 _____

method n°3 _____

➤ **Attrition bias about tumor growth curves**

η1. Number of animals assigned to each treatment arm (results section)

Exact Estimate Unknown
specify N° _____

η2. Are there animals assigned to each treatment arm and excluded from statistical analysis?

Yes No Unknown

η2 a. If yes, are reasons for exclusion reported? Yes No Not applicable

Specify reasons for exclusion _____

η3. Are there animals at risk progressively reported in the plot of tumor growth curves?

Yes No Not applicable

η4. Is the number of animals with right-censored data clearly reported?

Yes No

η5. For each animal, is it clearly reported which event determined the end of follow-up?

Yes No

- η6.** Is the length of follow-up clearly defined? Yes No
Specify definition _____
- **Miscellanea**
- θ1.** State any important concerns about statistical design and analysis not covered by other sections in the checklist _____
- θ2.** Was any author a member of a department of statistics or epidemiology? Yes No
- θ3.** Was mentioned the use of a statistical software for data analysis? Yes No
Specify statistical software _____

Appendix C

List of eligible clinical trials

- Annunziata CM, Walker AJ, Minasian L, Yu M, Kotz H, Wood BJ, Calvo K, Choyke P, Kimm D, Steinberg SM, Kohn EC. Vandetanib, designed to inhibit VEGFR2 and EGFR signaling, had no clinical activity as monotherapy for recurrent ovarian cancer and no detectable modulation of VEGFR2. *Clin Cancer Res*. 2010 Jan 15;16(2):664-72
- Aoki D, Katsumata N, Nakanishi T, Kigawa J, Fujiwara K, Takehara K, Kamiura S, Hiura M, Hatae M, Sugiyama T, Ochiai K, Noda K. A phase II clinical trial of topotecan in Japanese patients with relapsed ovarian carcinoma. *Jpn J Clin Oncol*. 2011 Mar;41(3):320-7
- Audeh MW, Carmichael J, Penson RT, Friedlander M, Powell B, Bell-McGuinn KM, Scott C, Weitzel JN, Oaknin A, Loman N, Lu K, Schmutzler RK, Matulonis U, Wickens M, Tutt A. Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial. *Lancet*. 2010 Jul 24;376(9737):245-51
- Baker AF, Roe DJ, Laughren C, Cohen JL, Wright HM, Clouser MC, Cui H, Alberts DS, Chambers SK. Investigation of bendamustine HCL in a phase 2 study in women with resistant ovarian cancer. *Invest New Drugs*. 2013 Feb;31(1):160-6
- Baumann KH, du Bois A, Meier W, Rau J, Wimberger P, Sehouli J, Kurzeder C, Hilpert F, Hasenburg A, Canzler U, Hanker LC, Hillemanns P, Richter B, Wollschlaeger K, Dewitz T, Bauerschlag D, Wagner U. A phase II trial (AGO 2.11) in platinum-resistant ovarian cancer: a randomized multicenter trial with sunitinib (SU11248) to evaluate dosage, schedule, tolerability, toxicity and effectiveness of a multitargeted receptor tyrosine kinase inhibitor monotherapy. *Ann Oncol*. 2012 Sep;23(9):2265-71
- Behbakht K, Sill MW, Darcy KM, Rubin SC, Mannel RS, Waggoner S, Schilder RJ, Cai KQ, Godwin AK, Alpaugh RK. Phase II trial of the mTOR inhibitor, temsirolimus and evaluation of circulating tumor cells and tumor biomarkers in persistent and recurrent epithelial ovarian and primary peritoneal malignancies: a Gynecologic Oncology Group study. *Gynecol Oncol*. 2011 Oct;123(1):19-26
- Bell-McGuinn KM¹, Konner JA, Tew WP, Hensley ML, Iasonos A, Charpentier E, Mironov S, Sabbatini P, Aghajanian C. A Phase 2, Single Arm Study of Iniparib in Patients With BRCA1 or BRCA2 Associated Advanced Epithelial Ovarian, Fallopian Tube, or Primary Peritoneal Cancer. *Int J Gynecol Cancer*. 2016 Feb;26(2):255-60

- Biagi JJ, Oza AM, Chalchal HI, Grimshaw R, Ellard SL, Lee U, Hirte H, Sederias J, Ivy SP, Eisenhauer EA. A phase II study of sunitinib in patients with recurrent epithelial ovarian and primary peritoneal carcinoma: an NCIC Clinical Trials Group Study. *Ann Oncol*. 2011 Feb;22(2):335-40
- Bodnar L, Górnas M, Szczylik C. Sorafenib as a third line therapy in patients with epithelial ovarian cancer or primary peritoneal cancer: a phase II study. *Gynecol Oncol*. 2011 Oct;123(1):33-6
- Brucker J, Mayer C, Gebauer G, Mallmann P, Belau AK, Schneeweiss A, Sohn C, Eichbaum M. Non-pegylated liposomal doxorubicin for patients with recurrent ovarian cancer: A multicentric phase II trial. *Oncol Lett*. 2016 Aug;12(2):1211-1215
- Burger RA, Deng W, Makker V, Collins Y, Gray H, Debernardo R, Martin LP, Aghajanian C. Phase II evaluation of dalantercept in the treatment of persistent or recurrent epithelial ovarian cancer: An NRG Oncology/Gynecologic Oncology Group study. *Gynecol Oncol*. 2018 Sep;150(3):466-470
- Campos SM, Penson RT, Matulonis U, Horowitz NS, Whalen C, Pereira L, Tyburski K, Roche M, Szymonifka J, Berlin S. A phase II trial of Sunitinib malate in recurrent and refractory ovarian, fallopian tube and peritoneal carcinoma. *Gynecol Oncol*. 2013 Feb;128(2):215-20
- Chan JK, Brady W, Monk BJ, Brown J, Shahin MS, Rose PG, Kim JH, Secord AA, Walker JL, Gershenson DM. A phase II evaluation of sunitinib in the treatment of persistent or recurrent clear cell ovarian carcinoma: An NRG Oncology/Gynecologic Oncology Group Study (GOG-254). *Gynecol Oncol*. 2018 Aug;150(2):247-252
- Coleman RL, Brady WE, McMeekin DS, Rose PG, Soper JT, Lentz SS, Hoffman JS, Shahin MS. A phase II evaluation of nanoparticle, albumin-bound (nab) paclitaxel in the treatment of recurrent or persistent platinum-resistant ovarian, fallopian tube, or primary peritoneal cancer: a Gynecologic Oncology Group study. *Gynecol Oncol*. 2011 Jul;122(1):111-5
- Coleman RL, Sill MW, Bell-McGuinn K, Aghajanian C, Gray HJ, Tewari KS, Rubin SC, Rutherford TJ, Chan JK, Chen A, Swisher EM. A phase II evaluation of the potent, highly selective PARP inhibitor veliparib in the treatment of persistent or recurrent epithelial ovarian, fallopian tube, or primary peritoneal cancer in patients who carry a germline BRCA1 or BRCA2 mutation - An NRG Oncology/Gynecologic Oncology Group study. *Gynecol Oncol*. 2015 Jun;137(3):386-91
- De Geest K, Blessing JA, Morris RT, Yamada SD, Monk BJ, Zweizig SL, Matei D, Muller CY, Richards WE. Phase II clinical trial of ixabepilone in patients with recurrent or persistent platinum- and taxane-resistant ovarian or primary peritoneal cancer: a gynecologic oncology group study. *J Clin Oncol*. 2010 Jan 1;28(1):149-53

- Diaz-Padilla I, Wilson MK, Clarke BA, Hirte HW, Welch SA, Mackay HJ, Biagi JJ, Reedijk M, Weberpals JJ, Fleming GF, Wang L, Liu G, Zhou C, Blattler C, Ivy SP, Oza AM. A phase II study of single-agent RO4929097, a gamma-secretase inhibitor of Notch signaling, in patients with recurrent platinum-resistant epithelial ovarian cancer: A study of the Princess Margaret, Chicago and California phase II consortia. *Gynecol Oncol*. 2015 May;137(2):216-22
- Drew Y, Ledermann J, Hall G, Rea D, Glasspool R, Highley M, Jayson G, Sludden J, Murray J, Jamieson D, Halford S, Acton G, Backholer Z, Mangano R, Boddy A, Curtin N, Plummer R. Phase 2 multicentre trial investigating intermittent and continuous dosing schedules of the poly(ADP-ribose) polymerase inhibitor rucaparib in germline BRCA mutation carriers with advanced ovarian and breast cancer. *Br J Cancer*. 2016 Mar 29;114(7):723-30
- Emons G, Gorchev G, Sehouli J, Wimberger P, Stähle A, Hanker L, Hilpert F, Sindermann H, Gründker C, Harter P. Efficacy and safety of AEZS-108 (INN: zoptarelin doxorubicin acetate) an LHRH agonist linked to doxorubicin in women with platinum refractory or resistant ovarian cancer expressing LHRH receptors: a multicenter phase II trial of the ago-study group (AGO GYN 5). *Gynecol Oncol*. 2014 Jun;133(3):427-32
- Emons G, Kurzeder C, Schmalfeldt B, Neuser P, de Gregorio N, Pfisterer J, Park-Simon TW, Mahner S, Schröder W, Lück HJ, Heubner ML, Hanker L, Thiel F, Hilpert F. Temsirolimus in women with platinum-refractory/resistant ovarian cancer or advanced/recurrent endometrial carcinoma. A phase II study of the AGO-study group (AGO-GYN8). *Gynecol Oncol*. 2016 Mar;140(3):450-6
- Escudier B, Faivre S, Van Cutsem E, Germann N, Pouget JC, Plummer R, Vergote I, Thistlethwaite F, Bjarnason GA, Jones R, Mackay H, Edeline J, Fartoux L, Hirte H, Oza A. A Phase II Multicentre, Open-Label, Proof-of-Concept Study of Tasquinimod in Hepatocellular, Ovarian, Renal Cell, and Gastric Cancers. *Target Oncol*. 2017 Oct;12(5):655-661
- Farley J, Brady WE, Vathipadiekal V, Lankes HA, Coleman R, Morgan MA, Mannel R, Yamada SD, Mutch D, Rodgers WH, Birrer M, Gershenson DM. Selumetinib in women with recurrent low-grade serous carcinoma of the ovary or peritoneum: an open-label, single-arm, phase 2 study. *Lancet Oncol*. 2013 Feb;14(2):134-40
- Friedlander M, Hancock KC, Rischin D, Messing MJ, Stringer CA, Matthys GM, Ma B, Hodge JP, Lager JJ. A Phase II, open-label study evaluating pazopanib in patients with recurrent ovarian cancer. *Gynecol Oncol*. 2010 Oct;119(1):32-7
- Garcia AA, Sill MW, Lankes HA, Godwin AK, Mannel RS, Armstrong DK, Carolla RL, Liepmann MK, Spirtos NM, Fischer EG, Leslie KK. A phase II evaluation of lapatinib in the treatment of persistent

or recurrent epithelial ovarian or primary peritoneal carcinoma: a gynecologic oncology group study. *Gynecol Oncol*. 2012 Mar;124(3):569-74

- Gelmon KA, Tischkowitz M, Mackay H, Swenerton K, Robidoux A, Tonkin K, Hirte H, Huntsman D, Clemons M, Gilks B, Yerushalmi R, Macpherson E, Carmichael J, Oza A. Olaparib in patients with recurrent high-grade serous or poorly differentiated ovarian carcinoma or triple-negative breast cancer: a phase 2, multicentre, open-label, non-randomised study. *Lancet Oncol*. 2011 Sep;12(9):852-61
- Gold MA, Brady WE, Lankes HA, Rose PG, Kelley JL, De Geest K, Crispens MA, Resnick KE, Howell SB. A phase II study of a urokinase-derived peptide (A6) in the treatment of persistent or recurrent epithelial ovarian, fallopian tube, or primary peritoneal carcinoma: a Gynecologic Oncology Group study. *Gynecol Oncol*. 2012 Jun;125(3):635-9
- Graziani SR, Vital CG, Morikawa AT, Van Eyll BM, Fernandes Junior HJ, Kalil Filho R, Maranhão RC. Phase II study of paclitaxel associated with lipid core nanoparticles (LDE) as third-line treatment of patients with epithelial ovarian carcinoma. *Med Oncol*. 2017 Sep;34(9):151
- Hasegawa K, Kagabu M, Mizuno M, Oda K, Aoki D, Mabuchi S, Kamiura S, Yamaguchi S, Aoki Y, Saito T, Yunokawa M, Takehara K, Okamoto A, Ochiai K, Kimura T. Phase II basket trial of perifosine monotherapy for recurrent gynecologic cancer with or without PIK3CA mutations. *Invest New Drugs*. 2017 Dec;35(6):800-812
- Hays JL, Kim G, Walker A, Annunziata CM, Lee JM, Squires J, Houston N, Steinberg SM, Kohn EC. A phase II clinical trial of polyethylene glycol-conjugated L-asparaginase in patients with advanced ovarian cancer: Early closure for safety. *Mol Clin Oncol*. 2013 May;1(3):565-569
- Hensley ML, Kravetz S, Jia X, Iasonos A, Tew W, Pereira L, Sabbatini P, Whalen C, Aghajanian CA, Zarwan C, Berlin S. Eribulin mesylate (halichondrin B analog E7389) in platinum-resistant and platinum-sensitive ovarian cancer: a 2-cohort, phase 2 study. *Cancer*. 2012 May 1;118(9):2403-10
- Herzog TJ, Sill MW, Walker JL, O'Malley D, Shahin M, DeGeest K, Weiner SA, Mutch D, DeBernardo RL, Lentz SS. A phase II study of two topotecan regimens evaluated in recurrent platinum-sensitive ovarian, fallopian tube or primary peritoneal cancer: a Gynecologic Oncology Group Study (GOG 146Q). *Gynecol Oncol*. 2011 Mar;120(3):454-8
- Hirte H, Lheureux S, Fleming GF, Sugimoto A, Morgan R, Biagi J, Wang L, McGill S, Ivy SP, Oza AM. A phase 2 study of cediranib in recurrent or persistent ovarian, peritoneal or fallopian tube

cancer: a trial of the Princess Margaret, Chicago and California Phase II Consortia. *Gynecol Oncol*. 2015 Jul;138(1):55-61

- Kaufman B, Shapira-Frommer R, Schmutzler RK, Audeh MW, Friedlander M, Balmaña J, Mitchell G, Fried G, Stemmer SM, Hubert A, Rosengarten O, Steiner M, Loman N, Bowen K, Fielding A, Domchek SM. Olaparib monotherapy in patients with advanced cancer and a germline BRCA1/2 mutation. *J Clin Oncol*. 2015 Jan 20;33(3):244-50
- Kaye SB, Lubinski J, Matulonis U, Ang JE, Gourley C, Karlan BY, Amnon A, Bell-McGuinn KM, Chen LM, Friedlander M, Safra T, Vergote I, Wickens M, Lowe ES, Carmichael J, Kaufman B. Phase II, open-label, randomized, multicenter study comparing the efficacy and safety of olaparib, a poly (ADP-ribose) polymerase inhibitor, and pegylated liposomal doxorubicin in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer. *J Clin Oncol*. 2012 Feb 1;30(4):372-9
- Konstantinopoulos PA, Brady WE, Farley J, Armstrong A, Uyar DS, Gershenson DM. Phase II study of single-agent cabozantinib in patients with recurrent clear cell ovarian, primary peritoneal or fallopian tube cancer (NRG-GY001). *Gynecol Oncol*. 2018 Jul;150(1):9-13
- Kristeleit R, Shapiro GI, Burris HA, Oza AM, LoRusso P, Patel MR, Domchek SM, Balmaña J, Drew Y, Chen LM, Safra T, Montes A, Giordano H, Maloney L, Goble S, Isaacson J, Xiao J, Borrow J, Rolfe L, Shapira-Frommer R. A Phase I-II Study of the Oral PARP Inhibitor Rucaparib in Patients with Germline BRCA1/2-Mutated Ovarian Carcinoma or Other Solid Tumors. *Clin Cancer Res*. 2017 Aug 1;23(15):4095-4106
- Lee JM, Nair J, Zimmer A, Lipkowitz S, Annunziata CM, Merino MJ, Swisher EM, Harrell MI, Trepel JB, Lee MJ, Bagheri MH, Botesteanu DA, Steinberg SM, Minasian L, Ekwede I, Kohn EC. Prexasertib, a cell cycle checkpoint kinase 1 and 2 inhibitor, in BRCA wild-type recurrent high-grade serous ovarian cancer: a first-in-class proof-of-concept phase 2 study. *Lancet Oncol*. 2018 Feb;19(2):207-215
- Lheureux S, Tinker A, Clarke B, Ghatage P, Welch S, Weberpals JI, Dhani NC, Butler MO, Tonkin K, Tan Q, Tan DSP, Brooks K, Ramsahai J, Wang L, Pham NA, Shaw PA, Tsao MS, Garg S, Stockley T, Oza AM. A Clinical and Molecular Phase II Trial of Oral ENMD-2076 in Ovarian Clear Cell Carcinoma (OCCC): A Study of the Princess Margaret Phase II Consortium. *Clin Cancer Res*. 2018 Dec 15;24(24):6168-6174
- Lorusso D, Scambia G, Pignata S, Sorio R, Amadio G, Lepori S, Mosconi A, Pisano C, Mangili G, Maltese G, Sabbatini R, Artioli G, Gamucci T, Di Napoli M, Capoluongo E, Ludovini V, Raspagliesi F, Ferrandina G. Prospective phase II trial of trabectedin in BRCA-mutated and/or BRCAness

phenotype recurrent ovarian cancer patients: the MITO 15 trial. *Ann Oncol*. 2016 Mar;27(3):487-93

- Mackay HJ, Hirte H, Colgan T, Covens A, MacAlpine K, Greci P, Wang L, Mason J, Pham PA, Tsao MS, Pan J, Zwiebel J, Oza AM. Phase II trial of the histone deacetylase inhibitor belinostat in women with platinum resistant epithelial ovarian cancer and micropapillary (LMP) ovarian tumours. *Eur J Cancer*. 2010 Jun;46(9):1573-9
- Matei D, Sill MW, Lankes HA, DeGeest K, Bristow RE, Mutch D, Yamada SD, Cohn D, Calvert V, Farley J, Petricoin EF, Birrer MJ. Activity of sorafenib in recurrent ovarian cancer and primary peritoneal carcinomatosis: a gynecologic oncology group trial. *J Clin Oncol*. 2011 Jan 1;29(1):69-75
- Matulonis UA, Lee J, Lasonde B, Tew WP, Yehwalashet A, Matei D, Behbakht K, Grothusen J, Fleming G, Lee NK, Arnott J, Bray MR, Fletcher G, Brokx RD, Castonguay V, Mackay H, Sidor CF, Oza AM. ENMD-2076, an oral inhibitor of angiogenic and proliferation kinases, has activity in recurrent, platinum resistant ovarian cancer. *Eur J Cancer*. 2013 Jan;49(1):121-3
- Matulonis UA, Sharma S, Ghamande S, Gordon MS, Del Prete SA, Ray-Coquard I, Kutarska E, Liu H, Fingert H, Zhou X, Danaee H, Schilder RJ. Phase II study of MLN8237 (alisertib), an investigational Aurora A kinase inhibitor, in patients with platinum-resistant or -refractory epithelial ovarian, fallopian tube, or primary peritoneal carcinoma. *Gynecol Oncol*. 2012 Oct;127(1):63-9
- Matulonis UA, Sill MW, Makker V, Mutch DG, Carlson JW, Darus CJ, Mannel RS, Bender DP, Crane EK, Aghajanian C. A randomized phase II study of cabozantinib versus weekly paclitaxel in the treatment of persistent or recurrent epithelial ovarian, fallopian tube or primary peritoneal cancer: An NRG Oncology/Gynecologic Oncology Group study. *Gynecol Oncol*. 2019 Mar;152(3):548-553
- Miao M, Deng G, Luo S, Zhou J, Chen L, Yang J, He J, Li J, Yao J, Tan S, Tang J. A phase II study of apatinib in patients with recurrent epithelial ovarian cancer. *Gynecol Oncol*. 2018 Feb;148(2):286-290
- Noguera IR, Sun CC, Broaddus RR, Branham D, Levenback CF, Ramirez PT, Sood AK, Coleman RL, Gershenson DM. Phase II trial of imatinib mesylate in patients with recurrent platinum- and taxane-resistant low-grade serous carcinoma of the ovary, peritoneum, or fallopian tube. *Gynecol Oncol*. 2012 Jun;125(3):640-5

- Noonan AM, Bunch KP, Chen JQ, Herrmann MA, Lee JM, Kohn EC, O'Sullivan CC, Jordan E, Houston N, Takebe N, Kinders RJ, Cao L, Peer CJ, Figg WD, Annunziata CM. Pharmacodynamic markers and clinical results from the phase 2 study of the SMAC mimetic birinapant in women with relapsed platinum-resistant or -refractory epithelial ovarian cancer. *Cancer*. 2016 Feb 15;122(4):588-597
- Pecorelli S, Ray-Coquard I, Tredan O, Colombo N, Parma G, Tisi G, Katsaròs D, Lhommé C, Lissoni AA, Vermorken JB, du Bois A, Poveda A, Frigerio L, Barbieri P, Carminati P, Brienza S, Guastalla JP. Phase II of oral gimatecan in patients with recurrent epithelial ovarian, fallopian tube or peritoneal cancer, previously treated with platinum and taxanes. *Ann Oncol*. 2010 Apr;21(4):759-65
- Pignata S, Amant F, Scambia G, Sorio R, Breda E, Rasch W, Hernes K, Pisano C, Leunen K, Lorusso D, Cannella L, Vergote I. A phase I-II study of elacytarabine (CP-4055) in the treatment of patients with ovarian cancer resistant or refractory to platinum therapy. *Cancer Chemother Pharmacol*. 2011 Nov;68(5):1347-53
- Poveda A, Del Campo JM, Ray-Coquard I, Alexandre J, Provansal M, Guerra Alía EM, Casado A, Gonzalez-Martin A, Fernández C, Rodriguez I, Soto A, Kahatt C, Fernández Teruel C, Galmarini CM, Pérez de la Haza A, Bohan P, Berton-Rigaud D. Phase II randomized study of PM01183 versus topotecan in patients with platinum-resistant/refractory advanced ovarian cancer. *Ann Oncol*. 2017 Jun 1;28(6):1280-1287
- Pujade-Lauraine E, Selle F, Weber B, Ray-Coquard IL, Vergote I, Sufliarsky J, Del Campo JM, Lortholary A, Lesoin A, Follana P, Freyer G, Pardo B, Vidal L, Tholander B, Gladieff L, Sassi M, Garin-Chesa P, Nazabadioko S, Marzin K, Pilz K, Joly F. Volasertib Versus Chemotherapy in Platinum-Resistant or -Refractory Ovarian Cancer: A Randomized Phase II Groupe des Investigateurs Nationaux pour l'Etude des Cancers de l'Ovaire Study. *J Clin Oncol*. 2016 Mar 1;34(7):706-13
- Rosner B, Glynn RJ, Lee MLT. Incorporation of Clustering Effects for the Wilcoxon Rank Sum Test: A Large-Sample Approach. *Biometrics* 2003;59(4):1089-1098
- Rosner B, Grove JD. Use of the Mann-Whitney U-Test for Clustered Data. *Statist. Med*. 1999;18(11):1387-1400
- Rustin G, Reed N, Jayson GC, Ledermann JA, Adams M, Perren T, Poole C, Lind M, Persic M, Essapen S, Gore M, Calvert H, Stredder C, Wagner A, Giurescu M, Kaye S. A phase II trial evaluating

two schedules of sagopilone (ZK-EPO), a novel epothilone, in patients with platinum-resistant ovarian cancer. *Ann Oncol*. 2011 Nov;22(11):2411-6

- Schilder RJ, Blessing JA, Shahin MS, Miller DS, Tewari KS, Muller CY, Warshal DP, McMeekin S, Rotmensch J. A phase 2 evaluation of irofulven as second-line treatment of recurrent or persistent intermediately platinum-sensitive ovarian or primary peritoneal cancer: a Gynecologic Oncology Group trial. *Int J Gynecol Cancer*. 2010 Oct;20(7):1137-41
- Schilder RJ, Brady WE, Lankes HA, Fiorica JV, Shahin MS, Zhou XC, Mannel RS, Pathak HB, Hu W, Alpaugh RK, Sood AK, Godwin AK. Phase II evaluation of dasatinib in the treatment of recurrent or persistent epithelial ovarian or primary peritoneal carcinoma: a Gynecologic Oncology Group study. *Gynecol Oncol*. 2012 Oct;127(1):70-4
- Schilder RJ, Sill MW, Lankes HA, Gold MA, Mannel RS, Modesitt SC, Hanjani P, Bonebrake AJ, Sood AK, Godwin AK, Hu W, Alpaugh RK. A phase II evaluation of motesanib (AMG 706) in the treatment of persistent or recurrent ovarian, fallopian tube and primary peritoneal carcinomas: a Gynecologic Oncology Group study. *Gynecol Oncol*. 2013 Apr;129(1):86-91
- Schöffski P, Besse B, Gauler T, de Jonge MJ, Scambia G, Santoro A, Davite C, Jannuzzo MG, Petroccione A, Delord JP. Efficacy and safety of biweekly i.v. administrations of the Aurora kinase inhibitor danusertib hydrochloride in independent cohorts of patients with advanced or metastatic breast, ovarian, colorectal, pancreatic, small-cell and non-small-cell lung cancer: a multi-tumour, multi-institutional phase II study. *Ann Oncol*. 2015 Mar;26(3):598-607
- Schöffski P, Blay JY, De Greve J, Brain E, Machiels JP, Soria JC, Sleijfer S, Wolter P, Ray-Coquard I, Fontaine C, Munzert G, Fritsch H, Hanft G, Aerts C, Rapon J, Allgeier A, Bogaerts J, Lacombe D. Multicentric parallel phase II trial of the polo-like kinase 1 inhibitor BI 2536 in patients with advanced head and neck cancer, breast cancer, ovarian cancer, soft tissue sarcoma and melanoma. The first protocol of the European Organization for Research and Treatment of Cancer (EORTC) Network Of Core Institutes (NOCI). *Eur J Cancer*. 2010 Aug;46(12):2206-15
- Schwandt A, von Gruenigen VE, Wenham RM, Frasure H, Eaton S, Fusco N, Fu P, Wright JJ, Dowlati A, Waggoner S. Randomized phase II trial of sorafenib alone or in combination with carboplatin/paclitaxel in women with recurrent platinum sensitive epithelial ovarian, peritoneal, or fallopian tube cancer. *Invest New Drugs*. 2014 Aug;32(4):729-38
- Secord AA, McCollum M, Davidson BA, Broadwater G, Squatrito R, Havrilesky LJ, Gabel AC, Starr MD, Brady JC, Nixon AB, Duska LR. Phase II trial of nintedanib in patients with bevacizumab-

resistant recurrent epithelial ovarian, tubal, and peritoneal cancer. *Gynecol Oncol*. 2019 Jun;153(3):555-561

- Seetharamu N, Kim E, Hochster H, Martin F, Muggia F. Phase II study of liposomal cisplatin (SPI-77) in platinum-sensitive recurrences of ovarian cancer. *Anticancer Res*. 2010 Feb;30(2):541-5
- Sehouli J, Stengel D, Harter P, Kurzeder C, Belau A, Bogenrieder T, Markmann S, Mahner S, Mueller L, Lorenz R, Nugent A, Wilke J, Kuznik A, Doering G, Wischnik A, Sommer H, Meerpohl HG, Schroeder W, Lichtenegger W, Oskay-Oezcelik G. Topotecan Weekly Versus Conventional 5-Day Schedule in Patients With Platinum-Resistant Ovarian Cancer: a randomized multicenter phase II trial of the North-Eastern German Society of Gynecological Oncology Ovarian Cancer Study Group. *J Clin Oncol*. 2011 Jan 10;29(2):242-8
- Selle F, Sevin E, Ray-Coquard I, Mari V, Berton-Rigaud D, Favier L, Fabbro M, Lesoin A, Lortholary A, Pujade-Lauraine E. A phase II study of lenalidomide in platinum-sensitive recurrent ovarian carcinoma. *Ann Oncol*. 2014 Nov;25(11):2191-6
- Smit WM, Šufliarsky J, Werner TL, Dizon DS, Wagnerová M, Hirte HW, Spirtos NM, Oza A, Dirix L, El-Hashimy M, Acharyya S, Tan EY, Weber D, Schellens JHM. A Phase II Study of Patupilone (EPO906) in Patients With Platinum-Resistant or Platinum-Refractory Ovarian Cancer. *Clinical Ovarian and Other Gynecologic Cancer*. 2012;5(2): 53-9
- Steffensen KD, Adimi P, Jakobsen A. Veliparib Monotherapy to Patients With BRCA Germ Line Mutation and Platinum-Resistant or Partially Platinum-Sensitive Relapse of Epithelial Ovarian Cancer: A Phase I/II Study. *Int J Gynecol Cancer*. 2017 Nov;27(9):1842-1849
- Swisher EM, Lin KK, Oza AM, Scott CL, Giordano H, Sun J, Konecny GE, Coleman RL, Tinker AV, O'Malley DM, Kristeleit RS, Ma L, Bell-McGuinn KM, Brenton JD, Cragun JM, Oaknin A, Ray-Coquard I, Harrell MI7, Mann E, Kaufmann SH, Floquet A, Leary A, Harding TC, Goble S, Maloney L, Isaacson J, Allen AR, Rolfe L, Yelensky R, Raponi M, McNeish IA. Rucaparib in relapsed, platinum-sensitive high-grade ovarian carcinoma (ARIEL2 Part 1): an international, multicentre, open-label, phase 2 trial. *Lancet Oncol*. 2017 Jan;18(1):75-87
- Tew WP, Colombo N, Ray-Coquard I, Del Campo JM, Oza A, Pereira D, Mammoliti S, Matei D, Scambia G, Tonkin K, Shun Z, Sternas L, Spriggs DR. Intravenous aflibercept in patients with platinum-resistant, advanced ovarian cancer: results of a randomized, double-blind, phase 2, parallel-arm study. *Cancer*. 2014 Feb 1;120(3):335-43
- Usha L, Sill MW, Darcy KM, Benbrook DM, Hurteau JA, Michelin DP, Mannel RS, Hanjani P, De Geest K, Godwin AK. A Gynecologic Oncology Group phase II trial of the protein kinase C-beta

inhibitor, enzastaurin and evaluation of markers with potential predictive and prognostic value in persistent or recurrent epithelial ovarian and primary peritoneal malignancies. *Gynecol Oncol*. 2011 Jun 1;121(3):455-61

- Vergote IB, Garcia A, Michal J, Pippitt C, Bendell J, Spitz D, Reed N, Dark G, Fracasso PM, Ibrahim EN, Armenio VA, Duska L, Poole C, Gennigens C, Dirix LY, Leung AC, Zhao C, Soufi-Mahjoubi R, Rustin G. Randomized multicenter phase II trial comparing two schedules of etirinotecan pegol (NKTR-102) in women with recurrent platinum-resistant/refractory epithelial ovarian cancer. *J Clin Oncol*. 2013 Nov 10;31(32):4060-6.
- Vergote IB, Smith DC, Berger R, Kurzrock R, Vogelzang NJ, Sella A, Wheler J, Lee Y, Foster PG, Weitzman R, Buckanovich RJ. A phase 2 randomised discontinuation trial of cabozantinib in patients with ovarian carcinoma. *Eur J Cancer*. 2017 Sep;83:229-236

Appendix D

List of eligible preclinical *in vivo* experiments

- Alvero AB, Brown D, Montagna M, Matthews M, Mor G. Phenoxodiol-Topotecan co-administration exhibit significant anti-tumor activity without major adverse side effects. *Cancer Biol Ther.* 2007 Apr;6(4):612-7
- Asano M, Matsui J, Towle MJ, Wu J, McGonigle S, DE Boisferon MH, Uenaka T, Nomoto K, Littlefield BA. Broad-spectrum Preclinical Antitumor Activity of Eribulin (Halaven®): Combination with Anticancer Agents of Differing Mechanisms. *Anticancer Res.* 2018 Jun;38(6):3375-3385
- Bajrami I, Frankum JR, Konde A, Miller RE, Rehman FL, Brough R, Campbell J, Sims D, Rafiq R, Hooper S, Chen L, Kozarewa I, Assiotis I, Fenwick K, Natrajan R, Lord CJ, Ashworth A. Genome-wide profiling of genetic synthetic lethality identifies CDK12 as a novel determinant of PARP1/2 inhibitor sensitivity. *Cancer Res.* 2014 Jan 1;74(1):287-97
- Bani M, Decio A, Giavazzi R, Ghilardi C. Contribution of tumor endothelial cells to drug resistance: anti-angiogenic tyrosine kinase inhibitors act as p-glycoprotein antagonists. *Angiogenesis.* 2017 May;20(2):233-241
- Bartholomeusz C, Oishi T, Saso H, Akar U, Liu P, Kondo K, Kazansky A, Krishnamurthy S, Lee J, Esteva FJ, Kigawa J, Ueno NT. MEK1/2 inhibitor selumetinib (AZD6244) inhibits growth of ovarian clear cell carcinoma in a PEA-15-dependent manner in a mouse xenograft model. *Mol Cancer Ther.* 2012 Feb;11(2):360-9
- Bauerschlag DO, Schem C, Tiwari S, Egberts JH, Weigel MT, Kalthoff H, Jonat W, Maass N, Meinhold-Heerlein I. Sunitinib (SU11248) inhibits growth of human ovarian cancer in xenografted mice. *Anticancer Res.* 2010 Sep;30(9):3355-60
- Becker MA, Farzan T, Harrington SC, Krempski JW, Weroha SJ, Hou X, Kalli KR, Wong TW, Haluska P. Dual HER/VEGF receptor targeting inhibits *in vivo* ovarian cancer tumor growth. *Mol Cancer Ther.* 2013 Dec;12(12):2909-16
- Carlier C, Strese S, Viktorsson K, Velandar E, Nygren P, Uustalu M, Juntti T, Lewensohn R, Larsson R, Spira J, De Vlieghere E, Ceelen WP, Gullbo J. Preclinical activity of melflufen (J1) in ovarian cancer. *Oncotarget.* 2016 Sep 13;7(37):59322-59335
- Cesca M, Frapolli R, Berndt A, Scarlato V, Richter P, Kosmehl H, D'Incalci M, Ryan AJ, Giavazzi R. The effects of vandetanib on paclitaxel tumor distribution and antitumor activity in a xenograft model of human ovarian carcinoma. *Neoplasia.* 2009 Nov;11(11):1155-64

- Chau WK, Ip CK, Mak AS, Lai HC, Wong AS. c-Kit mediates chemoresistance and tumor-initiating capacity of ovarian cancer cells through activation of Wnt/ β -catenin-ATP-binding cassette G2 signaling. *Oncogene*. 2013 May 30;32(22):2767-81
- Colmegna B, Ubaldi S, Frapolli R, Licandro SA, Panini N, Galmarini CM, Badri N, Spanswick VJ, Bingham JP, Kiakos K, Erba E, Hartley JA, D'Incalci M. Increased sensitivity to platinum drugs of cancer cells with acquired resistance to trabectedin. *Br J Cancer*. 2015 Dec 22;113(12):1687-93
- De Cesare M, Pratesi G, Perego P, Carenini N, Tinelli S, Merlini L, Penco S, Pisano C, Bucci F, Vesci L, Pace S, Capocasa F, Carminati P, Zunino F. Potent antitumor activity and improved pharmacological profile of ST1481, a novel 7-substituted camptothecin. *Cancer Res*. 2001 Oct 1;61(19):7189-95
- De Cesare M, Pratesi G, Veneroni S, Bergottini R, Zunino F. Efficacy of the novel camptothecin gimatecan against orthotopic and metastatic human tumor xenograft models. *Clin Cancer Res*. 2004 Nov 1;10(21):7357-64
- Decio A, Cesca M, Bizzaro F, Porcu L, Bettolini R, Ubezio P, Taraboletti G, Belotti D, Giavazzi R. Cediranib combined with chemotherapy reduces tumor dissemination and prolongs the survival of mice bearing patient-derived ovarian cancer xenografts with different responsiveness to cisplatin. *Clin Exp Metastasis*. 2015 Oct;32(7):647-58
- Desai N, Trieu V, Yao Z, Louie L, Ci S, Yang A, Tao C, De T, Beals B, Dykes D, Noker P, Yao R, Labao E, Hawkins M, Soon-Shiong P. Increased antitumor activity, intratumor paclitaxel concentrations, and endothelial cell transport of cremophor-free, albumin-bound paclitaxel, ABI-007, compared with cremophor-based paclitaxel. *Clin Cancer Res*. 2006 Feb 15;12(4):1317-24
- Devapatla B, Sharma A, Woo S. CXCR2 Inhibition Combined with Sorafenib Improved Antitumor and Antiangiogenic Response in Preclinical Models of Ovarian Cancer. *PLoS One*. 2015 Sep 28;10(9):e0139237
- DeVorkin L, Hattersley M, Kim P, Ries J, Spowart J, Anglesio MS, Levi SM, Huntsman DG, Amaravadi RK, Winkler JD, Tinker AV, Lum JJ. Autophagy Inhibition Enhances Sunitinib Efficacy in Clear Cell Ovarian Carcinoma. *Mol Cancer Res*. 2017 Mar;15(3):250-258
- Dings RP, Van Laar ES, Webber J, Zhang Y, Griffin RJ, Waters SJ, MacDonald JR, Mayo KH. Ovarian tumor growth regression using a combination of vascular targeting agents anginex or topomimetic 0118 and the chemotherapeutic irifolven. *Cancer Lett*. 2008 Jul 8;265(2):270-80

- Do TV, Xiao F, Bickel LE, Klein-Szanto AJ, Pathak HB, Hua X, Howe C, O'Brien SW, Maglaty M, Ecsedy JA, Litwin S, Golemis EA, Schilder RJ, Godwin AK, Connolly DC. Aurora kinase A mediates epithelial ovarian cancer cell migration and adhesion. *Oncogene*. 2014 Jan 30;33(5):539-49
- Erriquez J, Olivero M, Mittica G, Scalzo MS, Vaira M, De Simone M, Ponzzone R, Katsaros D, Aglietta M, Calogero R, Di Renzo MF, Valabrega G. Xenopatients show the need for precision medicine approach to chemotherapy in ovarian cancer. *Oncotarget*. 2016 May 3;7(18):26181-91
- Feng Z, Zhao G, Yu L, Gough D, Howell SB. Preclinical efficacy studies of a novel nanoparticle-based formulation of paclitaxel that out-performs Abraxane. *Cancer Chemother Pharmacol*. 2010 Apr;65(5):923-30
- Frapolli R, Zucchetti M, Sessa C, Marsoni S, Viganò L, Locatelli A, Rulli E, Compagnoni A, Bello E, Pisano C, Carminati P, D'Incalci M. Clinical pharmacokinetics of the new oral camptothecin gimatecan: the inter-patient variability is related to alpha1-acid glycoprotein plasma levels. *Eur J Cancer*. 2010 Feb;46(3):505-16
- Gaborit N, Larbouret C, Vallaghe J, Peyrusson F, Bascoul-Mollevi C, Crapez E, Azria D, Chardès T, Poul MA, Mathis G, Bazin H, Pèlegri A. Time-resolved fluorescence resonance energy transfer (TR-FRET) to analyze the disruption of EGFR/HER2 dimers: a new method to evaluate the efficiency of targeted therapy using monoclonal antibodies. *J Biol Chem*. 2011 Apr 1;286(13):11337-45
- Gonzalez-Villasana V, Fuentes-Mattei E, Ivan C, Dalton HJ, Rodriguez-Aguayo C, Fernandez-de Thomas RJ, Aslan B, Del C Monroig P, Velazquez-Torres G, Previs RA, Pradeep S, Kahraman N, Wang H, Kanlikilicer P, Ozpolat B, Calin G, Sood AK, Lopez-Berestein G. Rac1/Pak1/p38/MMP-2 Axis Regulates Angiogenesis in Ovarian Cancer. *Clin Cancer Res*. 2015 May 1;21(9):2127-37
- Guichard S, Montazeri A, Chatelut E, Hennebelle I, Bugat R, Canal P. Schedule-dependent activity of topotecan in OVCAR-3 ovarian carcinoma xenograft: pharmacokinetic and pharmacodynamic evaluation. *Clin Cancer Res*. 2001 Oct;7(10):3222-8
- Haemmerle M, Bottsford-Miller J, Pradeep S, Taylor ML, Choi HJ, Hansen JM, Dalton HJ, Stone RL, Cho MS, Nick AM, Nagaraja AS, Gutschner T, Gharpure KM, Mangala LS, Rupaimoole R, Han HD, Zand B, Armaiz-Pena GN, Wu SY, Pecot CV, Burns AR, Lopez-Berestein G, Afshar-Kharghan V, Sood AK. FAK regulates platelet extravasation and tumor growth after antiangiogenic therapy withdrawal. *J Clin Invest*. 2016 May 2;126(5):1885-96

- Hashimoto K, Man S, Xu P, Cruz-Munoz W, Tang T, Kumar R, Kerbel RS. Potent preclinical impact of metronomic low-dose oral topotecan combined with the antiangiogenic drug pazopanib for the treatment of ovarian cancer. *Mol Cancer Ther.* 2010 Apr;9(4):996-1006
- Hennessy BT, Lu Y, Poradosu E, Yu Q, Yu S, Hall H, Carey MS, Ravoory M, Gonzalez-Angulo AM, Birch R, Henderson IC, Kundra V, Mills GB. Pharmacodynamic markers of perifosine efficacy. *Clin Cancer Res.* 2007 Dec 15;13(24):7421-31
- Hew KE, Miller PC, El-Ashry D, Sun J, Besser AH, Ince TA, Gu M, Wei Z, Zhang G, Brafford P, Gao W, Lu Y, Mills GB, Slingerland JM, Simpkins F. MAPK Activation Predicts Poor Outcome and the MEK Inhibitor, Selumetinib, Reverses Antiestrogen Resistance in ER-Positive High-Grade Serous Ovarian Cancer. *Clin Cancer Res.* 2016 Feb 15;22(4):935-47
- Hijaz M, Chhina J, Mert I, Taylor M, Dar S, Al-Wahab Z, Ali-Fehmi R, Buekers T, Munkarah AR, Rattan R. Preclinical evaluation of olaparib and metformin combination in BRCA1 wildtype ovarian cancer. *Gynecol Oncol.* 2016 Aug;142(2):323-31
- Hirai H, Sootome H, Nakatsuru Y, Miyama K, Taguchi S, Tsujioka K, Ueno Y, Hatch H, Majumder PK, Pan BS, Kotani H. MK-2206, an allosteric Akt inhibitor, enhances antitumor efficacy by standard chemotherapeutic agents or molecular targeted drugs in vitro and in vivo. *Mol Cancer Ther.* 2010 Jul;9(7):1956-67
- Hoch U, Staschen CM, Johnson RK, Eldon MA. Nonclinical pharmacokinetics and activity of etirinotecan pegol (NKTR-102), a long-acting topoisomerase 1 inhibitor, in multiple cancer models. *Cancer Chemother Pharmacol.* 2014 Dec;74(6):1125-37
- Huang J, Hu W, Hu L, Previs RA, Dalton HJ, Yang XY, Sun Y, McGuire M, Rupaimoole R1, Nagaraja AS, Kang Y, Liu T, Nick AM, Jennings NB, Coleman RL, Jaffe RB, Sood AK. DII4 Inhibition plus Aflibercept Markedly Reduces Ovarian Tumor Growth. *Mol Cancer Ther.* 2016 Jun;15(6):1344-52
- Kim H, George E, Ragland R, Rafail S, Zhang R, Krepler C, Morgan M, Herlyn M, Brown E, Simpkins F. Targeting the ATR/CHK1 Axis with PARP Inhibition Results in Tumor Regression in BRCA-Mutant Ovarian Cancer Models. *Clin Cancer Res.* 2017 Jun 15;23(12):3097-3108
- Kondrashova O, Topp M, Nesic K, Lieschke E, Ho GY, Harrell MI, Zapparoli GV, Hadley A, Holian R, Boehm E, Heong V, Sanij E, Pearson RB, Krais JJ, Johnson N, McNally O, Ananda S, Alsop K, Hutt KJ, Kaufmann SH, Lin KK, Harding TC, Traficante N; Australian Ovarian Cancer Study (AOCS), deFazio A, McNeish IA, Bowtell DD, Swisher EM, Dobrovic A, Wakefield MJ, Scott CL. Methylation of all BRCA1 copies predicts response to the PARP inhibitor rucaparib in ovarian carcinoma. *Nat Commun.* 2018 Sep 28;9(1):3970

- Konstantinopoulos PA, Wilson AJ, Saskowski J, Wass E, Khabele D. Suberoylanilide hydroxamic acid (SAHA) enhances olaparib activity by targeting homologous recombination DNA repair in ovarian cancer. *Gynecol Oncol*. 2014 Jun;133(3):599-606
- Le XF, Mao W, He G, Claret FX, Xia W, Ahmed AA, Hung MC, Siddik ZH, Bast RC Jr. The role of p27(Kip1) in dasatinib-enhanced paclitaxel cytotoxicity in human ovarian cancer cells. *J Natl Cancer Inst*. 2011 Sep 21;103(18):1403-22
- Le XF, Mao W, Lu Z, Carter BZ, Bast RC Jr. Dasatinib induces autophagic cell death in human ovarian cancer. *Cancer*. 2010 Nov 1;116(21):4980-90
- Lee FY, Borzilleri R, Fairchild CR, Kim SH, Long BH, Reventos-Suarez C, Vite GD, Rose WC, Kramer RA. BMS-247550: a novel epothilone analog with a mode of action similar to paclitaxel but possessing superior antitumor efficacy. *Clin Cancer Res*. 2001 May;7(5):1429-37
- Li Y, Luo K, Yin Y, Wu C, Deng M, Li L, Chen Y, Nowsheen S, Lou Z, Yuan J. USP13 regulates the RAP80-BRCA1 complex dependent DNA damage response. *Nat Commun*. 2017 Jun 1;8:15752
- Lin ZP, Zhu YL, Lo YC, Moscarelli J, Xiong A, Korayem Y, Huang PH, Giri S, LoRusso P, Ratner ES. Combination of triapine, olaparib, and cediranib suppresses progression of BRCA-wild type and PARP inhibitor-resistant epithelial ovarian cancer. *PLoS One*. 2018 Nov 16;13(11):e0207399
- Liu G, Yang D, Rupaimoole R, Pecot CV, Sun Y, Mangala LS, Li X, Ji P, Cogdell D, Hu L, Wang Y, Rodriguez-Aguayo C, Lopez-Berestein G, Shmulevich I, De Cecco L, Chen K, Mezzanzanica D, Xue F, Sood AK, Zhang W. Augmentation of response to chemotherapy by microRNA-506 through regulation of RAD51 in serous ovarian cancers. *J Natl Cancer Inst*. 2015 May 20;107(7). pii: djv108
- Mabuchi S, Hisamatsu T, Kawase C, Hayashi M, Sawada K, Mimura K, Takahashi K, Takahashi T, Kurachi H, Kimura T. The activity of trabectedin as a single agent or in combination with everolimus for clear cell carcinoma of the ovary. *Clin Cancer Res*. 2011 Jul 1;17(13):4462-73
- Matsumura N, Mandai M, Okamoto T, Yamaguchi K, Yamamura S, Oura T, Baba T, Hamanishi J, Kang HS, Matsui S, Mori S, Murphy SK, Konishi I. Sorafenib efficacy in ovarian clear cell carcinoma revealed by transcriptome profiling. *Cancer Sci*. 2010 Dec;101(12):2658-63
- Matsuo K, Nishimura M, Bottsford-Miller JN, Huang J, Komurov K, Armaiz-Pena GN, Shahzad MM, Stone RL, Roh JW, Sanguino AM, Lu C, Im DD, Rosenshien NB, Sakakibara A, Nagano T, Yamasaki M, Enomoto T, Kimura T, Ram PT, Schmeler KM, Gallick GE, Wong KK, Frumovitz M, Sood AK. Targeting SRC in mucinous ovarian carcinoma. *Clin Cancer Res*. 2011 Aug 15;17(16):5367-78
- Miller RE, Brough R, Bajrami I, Williamson CT, McDade S, Campbell J, Kigozi A, Rafiq R, Pemberton H, Natrajan R, Joel J, Astley H, Mahoney C, Moore JD, Torrance C, Gordan JD, Webber JT, Levin

- RS, Shokat KM, Bandyopadhyay S, Lord CJ, Ashworth A. Synthetic Lethal Targeting of ARID1A-Mutant Ovarian Clear Cell Tumors with Dasatinib. *Mol Cancer Ther*. 2016 Jul;15(7):1472-84
- Miyake T, Pradeep S, Wu SY, Rupaimoole R1, Zand B, Wen Y, Gharpure KM, Nagaraja AS, Hu W, Cho MS, Dalton HJ, Previs RA, Taylor ML, Hisamatsu T, Kang Y, Liu T, Shacham S, McCauley D, Hawke DH, Wiktorowicz JE, Coleman RL, Sood AK. XPO1/CRM1 Inhibition Causes Antitumor Effects by Mitochondrial Accumulation of eIF5A. *Clin Cancer Res*. 2015 Jul 15;21(14):3286-97
 - Nagengast WB, Lub-de Hooje MN, Oosting SF, den Dunnen WF, Warnders FJ, Brouwers AH, de Jong JR, Price PM, Hollema H, Hospers GA, Elsinga PH, Hesselink JW, Gietema JA, de Vries EG. VEGF-PET imaging is a noninvasive biomarker showing differential changes in the tumor during sunitinib treatment. *Cancer Res*. 2011 Jan 1;71(1):143-53
 - Nicoletto MO, Padrini R, Palumbo M, Ziade A, Ragazzi RS, Pratesi G, Artioli G, De Cesare M, Zunino F. Intralesional topotecan in advanced ovarian cancer: a clinical report, based on a preclinical study. *Oncol Rep*. 2002 Nov-Dec;9(6):1351-4
 - Merritt WM, Nick AM, Carroll AR, Lu C, Matsuo K, Dumble M, Jennings N, Zhang S, Lin YG, Spannuth WA, Kamat AA, Stone RL, Shahzad MM, Coleman RL, Kumar R, Sood AK. Bridging the gap between cytotoxic and biologic therapy with metronomic topotecan and pazopanib in ovarian cancer. *Mol Cancer Ther*. 2010 Apr;9(4):985-95
 - Patankar NA, Waterhouse D, Strutt D, Anantha M, Bally MB. Topophore C: a liposomal nanoparticle formulation of topotecan for treatment of ovarian cancer. *Invest New Drugs*. 2013 Feb;31(1):46-58
 - Plumb JA, Finn PW, Williams RJ, Bandara MJ, Romero MR, Watkins CJ, La Thangue NB, Brown R. Pharmacodynamic response and inhibition of growth of human tumor xenografts by the novel histone deacetylase inhibitor PXD101. *Mol Cancer Ther*. 2003 Aug;2(8):721-8
 - Pratesi G, De Cesare M, Carenini N, Perego P, Righetti SC, Cucco C, Merlini L, Pisano C, Penco S, Carminati P, Vesci L, Zunino F. Pattern of antitumor activity of a novel camptothecin, ST1481, in a large panel of human tumor xenografts. *Clin Cancer Res*. 2002 Dec;8(12):3904-9
 - Previs RA, Armaiz-Pena GN, Lin YG, Davis AN, Pradeep S, Dalton HJ, Hansen JM, Merritt WM, Nick AM, Langley RR, Coleman RL, Sood AK. Dual Metronomic Chemotherapy with Nab-Paclitaxel and Topotecan Has Potent Antiangiogenic Activity in Ovarian Cancer. *Mol Cancer Ther*. 2015 Dec;14(12):2677-86

- Qian X, LaRoche WJ, Ara G, Wu F, Petersen KD, Thougard A, Sehested M, Lichenstein HS, Jeffers M. Activity of PXD101, a histone deacetylase inhibitor, in preclinical ovarian cancer studies. *Mol Cancer Ther.* 2006 Aug;5(8):2086-95
- Romanelli S, Perego P, Pratesi G, Carenini N, Tortoreto M, Zunino F. In vitro and in vivo interaction between cisplatin and topotecan in ovarian carcinoma systems. *Cancer Chemother Pharmacol.* 1998;41(5):385-90
- Sasano T, Mabuchi S, Kuroda H, Kawano M, Matsumoto Y, Takahashi R, Hisamatsu T, Sawada K, Hashimoto K, Isobe A, Testa JR, Kimura T. Preclinical Efficacy for AKT Targeting in Clear Cell Carcinoma of the Ovary. *Mol Cancer Res.* 2015 Apr;13(4):795-806
- Saucier JM, Yu J, Gaikwad A, Coleman RL, Wolf JK, Smith JA. Determination of the optimal combination chemotherapy regimen for treatment of platinum-resistant ovarian cancer in nude mouse model. *J Oncol Pharm Pract.* 2007 Mar;13(1):39-45
- Shi W, Siemann DW. Targeting the tumor vasculature: enhancing antitumor efficacy through combination treatment with ZD6126 and ZD6474. *In Vivo.* 2005 Nov-Dec;19(6):1045-50
- Stany MP, Vathipadiekal V, Ozbun L, Stone RL, Mok SC, Xue H, Kagami T, Wang Y, McAlpine JN, Bowtell D, Gout PW, Miller DM, Gilks CB, Huntsman DG, Ellard SL, Wang YZ, Vivas-Mejia P, Lopez-Berestein G, Sood AK, Birrer MJ. Identification of novel therapeutic targets in microdissected clear cell ovarian cancers. *PLoS One.* 2011;6(7):e21121
- Steele N, Finn P, Brown R, Plumb JA. Combined inhibition of DNA methylation and histone acetylation enhances gene re-expression and drug sensitivity in vivo. *Br J Cancer.* 2009 Mar 10;100(5):758-63
- Sui H, Shi C, Yan Z, Li H. Combination of erlotinib and a PARP inhibitor inhibits growth of A2780 tumor xenografts due to increased autophagy. *Drug Des Devel Ther.* 2015 Jun 22;9:3183-90
- Sun C, Li N, Yang Z, Zhou B, He Y, Weng D, Fang Y, Wu P, Chen P, Yang X, Ma D, Zhou J, Chen G. miR-9 regulation of BRCA1 and ovarian cancer sensitivity to cisplatin and PARP inhibition. *J Natl Cancer Inst.* 2013 Nov 20;105(22):1750-8
- Towle MJ, Salvato KA, Budrow J, Wels BF, Kuznetsov G, Aalfs KK, Welsh S, Zheng W, Seletsky BM, Palme MH, Habgood GJ, Singer LA, Dipietro LV, Wang Y, Chen JJ, Quincy DA, Davis A, Yoshimatsu K, Kishi Y, Yu MJ, Littlefield BA. In vitro and in vivo anticancer activities of synthetic macrocyclic ketone analogues of halichondrin B. *Cancer Res.* 2001 Feb 1;61(3):1013-21
- Tsunetoh S, Terai Y, Sasaki H, Tanabe A, Tanaka Y, Sekijima T, Fujioka S, Kawaguchi H, Kanemura M, Yamashita Y, Ohmichi M. Topotecan as a molecular targeting agent which blocks the Akt and

VEGF cascade in platinum-resistant ovarian cancers. *Cancer Biol Ther*. 2010 Dec 1;10(11):1137-46

- van Laar ES, Izbicka E, Weitman S, Medina-Gundrum L, Macdonald JR, Waters SJ. Antitumor activity of irifolven against human ovarian cancer cell lines, human tumor colony-forming units, and xenografts. *Int J Gynecol Cancer*. 2004 Sep-Oct;14(5):824-31
- van Hattum AH, Pinedo HM, Schlüper HM, Erkelens CA, Tohgo A, Boven E. The activity profile of the hexacyclic camptothecin derivative DX-8951f in experimental human colon cancer and ovarian cancer. *Biochem Pharmacol*. 2002 Oct 15;64(8):1267-77
- Wang D, Wang M, Jiang N, Zhang Y, Bian X, Wang X, Roberts TM, Zhao JJ, Liu P, Cheng H. Effective use of PI3K inhibitor BKM120 and PARP inhibitor Olaparib to treat PIK3CA mutant ovarian cancer. *Oncotarget*. 2016 Mar 15;7(11):13153-66
- Wedge SR, Kendrew J, Hennequin LF, Valentine PJ, Barry ST, Brave SR, Smith NR, James NH, Dukes M, Curwen JO, Chester R, Jackson JA, Boffey SJ, Kilburn LL, Barnett S, Richmond GH, Wadsworth PF, Walker M, Bigley AL, Taylor ST, Cooper L, Beck S, Jürgensmeier JM, Ogilvie DJ. AZD2171: a highly potent, orally bioavailable, vascular endothelial growth factor receptor-2 tyrosine kinase inhibitor for the treatment of cancer. *Cancer Res*. 2005 May 15;65(10):4389-400
- Wedge SR, Ogilvie DJ, Dukes M, Kendrew J, Chester R, Jackson JA, Boffey SJ, Valentine PJ, Curwen JO, Musgrove HL, Graham GA, Hughes GD, Thomas AP, Stokes ES, Curry B, Richmond GH, Wadsworth PF, Bigley AL, Hennequin LF. ZD6474 inhibits vascular endothelial growth factor signaling, angiogenesis, and tumor growth following oral administration. *Cancer Res*. 2002 Aug 15;62(16):4645-55
- Williams J, Lucas PC, Griffith KA, Choi M, Fogoros S, Hu YY, Liu JR. Expression of Bcl-xL in ovarian carcinoma is associated with chemoresistance and recurrent disease. *Gynecol Oncol*. 2005 Feb;96(2):287-95
- Wilson AJ, Stubbs M, Liu P, Ruggeri B, Khabele D. The BET inhibitor INCB054329 reduces homologous recombination efficiency and augments PARP inhibitor activity in ovarian cancer. *Gynecol Oncol*. 2018 Jun;149(3):575-584
- Yi XF, Fan SM, Yao M, Feng YJ. Comparison of efficacy and toxicity profile between intraperitoneal and intravenous topotecan in human ovarian cancer xenografts. *Beijing Da Xue Xue Bao Yi Xue Ban*. 2006 Feb 18;38(1):88-91

- Zhang C, Zhou SS, Li XR, Wang BM, Lin NM, Feng LY, Zhang DY, Zhang LH, Wang JB, Pan JP. Enhanced antitumor activity by the combination of dasatinib and combretastatin A-4 in vitro and in vivo. *Oncol Rep.* 2013 Jun;29(6):2275-82
- Zucchetti M, Bonezzi K, Frapolli R, Sala F, Borsotti P, Zangarini M, Cvitkovic E, Noel K, Ubezio P, Giavazzi R, D'Incalci M, Taraboletti G. Pharmacokinetics and antineoplastic activity of galectin-1-targeting OTX008 in combination with sunitinib. *Cancer Chemother Pharmacol.* 2013 Oct;72(4):879-87

Appendix E

SAS MACRO programs

```
* ----- *
* Program.....: POWER.SAS *
* Scope.....: Calculating asymptotic power of the family of statistical tests *
* Version.....: 1.0 *
* Author.....: Luca Porcu *
* Data created: 20AUG2019 *
* Project: new statistical framework to analyze tumor growth curves *
* Warning: slopes in different intervals in the same animal are distributed independently. *
*           This assumption does not usually hold for real cases *
* Example: %POWER(delta= 0.04, *
*           sigma= 0.06, *
*           m= 10, *
*           n= 10, *
*           k= 5) *
* ----- *

%MACRO POWER(
delta=,                               /*Absolute effect size*/
sigma=,                               /*Slopes' standard deviation*/
m=,                                   /*Number of animals in the active arm*/
n=,                                   /*Number of animals in the control arm*/
k=,                                   /*Number of time intervals*/
alpha=                                /*Type I error*/
) ;
Data power;
    delta = &delta.;
    sigma = &sigma.;
    m=&m.;
    n=&n.;
    k=&k.;
```

*** Expectation and variance of the Mann-Whitney statistics under the null hypothesis. No ties are present. Slopes in different intervals in the same animal are assumed to be distributed independently;

$E0 = k * (1/2) * (m * n);$

$Var0 = k * (1/12) * (m * n) * (m + n + 1);$

*** Expectation and variance of the Mann-Whitney statistics under the alternative hypothesis (i.e. additive treatment effect). No ties are present. Slopes in different intervals in the same animal are assumed to be distributed independently. Refer to the Lehmann's book (2006), pp. 71-72 for proofs of the following formulas;

$ES = \text{delta} / \text{sigma};$

$p1 = \text{probnorm}(ES / \text{sqrt}(2));$

*** Because of the symmetry of the normal distribution p2 and p3 are equal;

$p2 = \text{probbnrm}(ES / \text{sqrt}(2), ES / \text{sqrt}(2), 0.5);$

$p3 = \text{probbnrm}(ES / \text{sqrt}(2), ES / \text{sqrt}(2), 0.5);$

$E1 = k * (m * n) * p1;$

$Var1 = k * ((m * n) * p1 * (1 - p1) + m * n * (n - 1) * (p2 - p1)^2 + m * n * (m - 1) * (p3 - p1)^2);$

*** Asymptotic power calculation;

$\text{cut} = E0 + (\text{probit}(1 - (\alpha / 2))) * \text{sqrt}(Var0);$

*** A continuity correction was introduced in asymptotic power calculation;

$\text{power} = 1 - \text{probnorm}((\text{cut} - (0.5 + E1)) / \text{sqrt}(Var1));$

run;

Proc print Data=power;

Run;

%MEND;


```

* ----- *
* Program.....: MW_PASYMP.SAS *
* Scope.....: Computing asymptotic p-value *
* Version.....: 1.0 *
* Author.....: Luca Porcu *
* Data created: 22AUG2019 *
* Project: new statistical framework to analyze tumor growth curves *
* Warning: slopes in different intervals in the same animal are distributed independently. *
*           This assumption does not usually hold for real cases *
* Example: %MW_PASYMP (treatment= Doxorubicin, *
*                   control= Vehicle, *
*                   interval= time1 time2 time3 time4 time5, *
*                   dSet= dSet) *
* ----- *

%MACRO MW_PASYMP(
treatment=,                               /*Treatment compared to the control arm*/
control=,                                 /*Control arm*/
interval=,                               /*Names of tumor volume variables, at different time points*/
dSet=                                     /*Name of the SAS dataset*/
);

*** The "MW" dataset is initialized. It will contain the Mann-Whitney statistics and its asymptotic
expectation and variance, at each time point;

Data MW;
    length time $ 250;
    time = "";
    if 0 = 1;
    run;

%LET count=1;
%DO %WHILE(%SCAN(&interval.,&count.,%str( )) ne %str( ));
%Let time=%SCAN(&interval.,&count.,%str( ));

*** Calculation of the number of slopes at each time interval in the active and control arm;

Proc sql noprint;

```

```

select count(&time.)
    into :Ctr
    from &dset.
    where arm = "&Control." and &time. ne .;
select count(&time.)
    into :Trt
    from &dset.
    where arm = "&Treatment." and &time. ne .;
quit;

```

*** Calculation of the Mann-Whitney statistics at each time interval. No ties are present;

```
%IF %EVAL(&Ctr. > 0 AND &Trt. > 0) %THEN %DO;
```

```
Proc transpose Data=&dset. out=MWset(drop= _name_) prefix=Trt;
```

```

var &time.;
where arm = "&Treatment." and &time. ne .;
run;

```

```
Proc sql;
```

```

create table auxMW1 as
select a.&time. as Ctr, b.*
    from &dset. as a, MWset as b
    where a.arm = "&Control.";
quit;

```

```
%LET dims = Trt%LEFT(&Trt.);
```

```
Data auxMW2 (drop= i);
```

```

set auxMW1;
array rango (1:&Trt.) Trt1-&dims.;
MW = 0;
do i=1 to &Trt.;
if Ctr > rango(i) then MW = MW + 1;
end;
run;

```

```
Proc sql;
```

```
create table auxMW3 as
```

```

select sum(MW) as MW
      from auxMW2;
quit;

```

*** Expectation and variance of the Mann-Whitney statistics under the null hypothesis at each time interval. No ties are present;

Data MW&time.;

```

length time $ 250;
set auxMW3;
time = "&time.";
expMW = (1/2)*(%EVAL(&Ctr.*&Trt.));
varMW = (1/12)*(%EVAL(&Ctr.*&Trt.)*%EVAL(&Ctr.+&Trt.+1));
run;

```

Data MW;

```

set MW MW&time.;
O_E = MW - expMW;
run;

```

%END;

%LET count=%EVAL(&count.+1);

%END;

Data _NULL_;

```

set MW end= last;
if last then call symput('clock',time);
run;

```

*** The "PValue" dataset is built. It contains the Mann-Whitney statistics and its asymptotic expectation and variance over all time points;

Proc sql;

```

create table auxPValue as
select sum(O_E) as OminusE, sum(MW) as MWobs, sum(expMW) as MWexp, sum(varMW)
as MWvariance
      from MW;
quit;

```

Data PValue;

```

set auxPValue;
time = "&clock.";
chi2 = (OminusE**2)/MWvariance;
PValue = 1-probchi(chi2,1);
format PValue pvalue5.3;
run;

Proc print Data=PValue;
label MWobs = "Observed Mann-Whitney statistics"
      MWexp = "Expected Mann-Whitney statistics"
      OminusE = "Observed minus expected Mann-Whitney statistics"
      MWvariance = "Asymptotic variance of the Mann-Whitney statistics"
      chi2 = "Chi-square (1 d.f.)"
      PValue = "Asymptotic p-value"
      time = "Last time point in which it was possible to compare control and treatment arms";
run;
%MEND;

```

```

* ----- *
* Program.....: MW_PEXACT.SAS *
* Scope.....: Calculating exact p-value *
* Version.....: 1.0 *
* Author.....: Luca Porcu *
* Data created: 30AUG2019 *
* Project: new statistical framework to analyze tumor growth curves *
* Example: %MW_PEXACT (treatment= Doxorubicin, *
*                control= Vehicle, *
*                interval= time1 time2 time3 time4 time5, *
*                dSet= dSet, *
*                nSimul= 500, *
*                seed= 10) *
* ----- *

%MACRO MW_PEXACT(
treatment=, /*Treatment compared to the control arm*/
control=, /*Control arm*/
interval=, /*Names of tumor volume variables, at different time points*/
dSet=, /*Name of the dataset*/
nSimul=, /*Number of simulations*/
seed= /*Random seed used to initialize SAS pseudorandom number generator*/
);

*** 1st step: calculation of the Mann-Whitney statistics over all time points;

Data MW;
    length time $ 250;
    time = "";
    if 0 = 1;
    run;

%LET count=1;
%LET countReal=1;
%DO %WHILE(%SCAN(&interval.,&count.,%str( )) ne %str( ));
%Let time=%SCAN(&interval.,&count.,%str( ));

```

```

Proc sql noprint;
    select count(&time.)
        into :Ctr
        from &dset.
        where arm = "&Control." and &time. ne .;
    select count(&time.)
        into :Trt
        from &dset.
        where arm = "&Treatment." and &time. ne .;
quit;

```

*** The number of observations assigned to the control arm are calculated. This number will be used for each resampling SAS dataset;

```
%IF &count.=1 %THEN %DO;
```

```

Proc sql noprint;
    select count(&time.)
        into :nObs
        from &dset.
        where arm = "&Control.";
        %END;

```

```
%IF %EVAL(&Ctr. > 0 AND &Trt. > 0) %THEN %DO;
```

```

Proc transpose Data=&dset. out=MWset(drop= _name_) prefix=Trt;
    var &time.;
    where arm = "&Treatment." and &time. ne .;
run;

```

```

Proc sql;
    create table auxMW1 as
    select a.&time. as Ctr, b.*
        from &dset. as a, MWset as b
        where a.arm = "&Control.";
quit;

```

```
%LET dimens = Trt%LEFT(&Trt.);
```

```
Data auxMW2 (drop= i);
```

```

set auxMW1;
array rango (1:&Trt.) Trt1-&dimens.;
MW = 0;
do i=1 to &Trt.;
if Ctr > rango(i) then MW = MW + 1;
end;
run;

Proc sql;
create table auxMW3 as
select sum(MW) as MW
from auxMW2;
quit;

Data MW&time.;
length time $ 250;
set auxMW3;
time = "&time.";
run;

Data MW;
set MW MW&time.;
run;

%LET countReal=%EVAL(&countReal.+1);
%END;
%LET count=%EVAL(&count.+1);
%END;

*** The Mann-Whitney statistics is calculated;

Proc sql noprint;
create table statObs as
select sum(MW) as statObs
from MW;
quit;

*** 2nd step: calculation of the exact Mann-Whitney distribution over all time points;

```

*** The "PValueExt" dataset is initialized. It will contain the exact distribution of the Mann-Whitney statistics under the null hypothesis;

```
Data PValueExt;
```

```
    if 0 = 1;
```

```
    run;
```

```
%DO j=1 %TO %EVAL(&nSimul.);
```

*** The "MW" dataset is initialized again. It will contain the observed Mann-Whitney statistics in each resampling dataset;

```
Data MW;
```

```
    if 0 = 1;
```

```
    run;
```

```
%Let root=%EVAL(&seed.+&j.);
```

```
proc surveyselect data=&dset.
```

```
    method=srs n=&nObs. out=SampleSRS noprint seed=&root.;
```

```
    where arm in ("&Control.", "&Treatment.");
```

```
run;
```

```
Proc sql noprint;
```

```
    select quote(subject)
```

```
        into :IDrandom separated by ','
```

```
        from SampleSRS;
```

```
    quit;
```

```
Data Sample;
```

```
    set &dset. (where= (arm in("&Control.", "&Treatment.")));
```

```
    if subject in(&IDrandom.) then arm = "&Control.";
```

```
    else arm = "&Treatment.";
```

```
    run;
```

```
%LET countExt=1;
```

```
%DO %WHILE(%EVAL(&countExt. < &countReal.));
```

```
%Let time=%SCAN(&interval., &countExt., %str( ));
```

```
Proc sql noprint;
```

```
    select count(&time.)
```

```
        into :Ctr
```



```

        from sample
        where arm = "&Control." and &time. ne .;
select count(&time.)
        into :Trt
        from sample
        where arm = "&Treatment." and &time. ne .;
quit;

%IF %EVAL(&Ctr. > 0 AND &Trt. > 0) %THEN %DO;
Proc transpose Data=sample out=MWset(drop= _name_) prefix=Trt;
    var &time.;
    where arm = "&Treatment." and &time. ne .;
run;

Proc sql;
    create table auxMW1 as
    select a.&time. as Ctr, b.*
        from sample as a, MWset as b
        where a.arm = "&Control.";
    quit;

%LET dimens = Trt%LEFT(&Trt.);

Data auxMW2 (drop= i);
    set auxMW1;
    array rango (1:&Trt.) Trt1-&dimens.;
    MW = 0;
    do i=1 to &Trt.;
        if Ctr > rango(i) then MW = MW + 1;
    end;
run;

Proc sql;
    create table auxMW3&j. as
        select sum(MW) as MW
            from auxMW2;
        quit;

```

```

Data MW&time.;
    length time $ 250;
    set auxMW3&j.;
    time = "&time.";
    run;

Data MW;
    set MW MW&time.;
    run;

%END;

%LET countExt=%EVAL(&countExt.+1);

%END;

Proc sql;
    create table auxPValueExt&j. as
    select sum(MW) as stat
        from MW;
    quit;

Data PValueExt;
    set PValueExt auxPValueExt&j.;
    run;

%END;

Proc sql;
    create table distrib1 as
    select a.stat, b.statObs
        from PValueExt as a, statObs as b;

*** Expected value and variance of exact Mann-Whitney distribution are stored in "distrib2" table;

    create table distrib2 as
        select mean(stat) as expected, var(stat) as varDistrib
            from distrib1;

    create table distrib3 as
    select a.*, b.expected, b.varDistrib
        from distrib1 as a, distrib2 as b;
    quit;

```

```

Data distrib;
    set distrib3;
    absDelta=abs(stat-expected);
    absObs=abs(statObs-expected);
    run;
*** 3rd step: exact p-value is calculated;
Proc sql;
    create table PValue as
    select (count(*)) as freq, (calculated freq) / &nSimul. as probMW format=pvalue5.3
        from distrib
        where absDelta >= absObs;
    quit;
Data exactDistrib;
    merge PValue distrib2;
    run;
Proc print Data=exactDistrib label;
var expected varDistrib probMW;
label expected = "Exact expected value"
    varDistrib = "Exact variance"
    probMW = "Exact p-value";
run;
%MEND;

```

```

* ----- *
* Program.....: MW_ESTIM.SAS *
* Scope.....: Estimating credibility intervals of the additive treatment effect *
* Version.....: 1.0 *
* Author.....: Luca Porcu *
* Data created: 02SEP2019 *
* Project: new statistical framework to analyze tumor growth curves *
* Example: %MW_ ESTIM (treatment= Doxorubicin, *
*               control= Vehicle, *
*               interval= time1 time2 time3 time4 time5, *
*               dSet= dSet, *
*               nSimul= 500, *
*               seed= 10, *
*               alpha= 0.05) *
* ----- *

%MACRO MW_ ESTIM(
treatment=,                /*Treatment to be compared to the control arm*/
control=,                  /*Control arm*/
interval=,                 /*Names of tumor volume variables, at different time points*/
dSet=,                     /*Name of the dataset*/
nSimul=,                   /*Number of simulations*/
seed=,                     /*Random seed used to initialize SAS pseudorandom number generator*/
alpha=                     /*Probability not covered by the credible interval*/
);

*** The "Delta" dataset is initialized. It will contain the set of slope differences between control and
treatment arms;

Data Delta;
    length time $ 250;
    time = "";
    if 0 = 1;
run;

```

*** The "MWexact" dataset is initialized. It will contain the exact distribution of the Mann-Whitney statistics under null hypothesis;

```
Data MWexact;
```

```
length time $ 250;
```

```
time = "";
```

```
if 0 = 1;
```

```
run;
```

```
%LET count=1;
```

```
%LET Pts=0;
```

```
%DO %WHILE(%SCAN(&interval.,&count.,%str( )) ne %str( ));
```

*** The number of observations at time &time. are calculated in control and treatment arms;

```
%Let time=%SCAN(&interval.,&count.,%str( ));
```

```
Proc sql noprint;
```

```
select count(&time.)
```

```
into :Ctr
```

```
from &dSet.
```

```
where arm = "&Control." and &time. ne .;
```

```
select count(&time.)
```

```
into :Trt
```

```
from &dset.
```

```
where arm = "&Treatment." and &time. ne .;
```

```
quit;
```

```
%IF %EVAL(&Trt. > &Pts.) %THEN %DO; %LET Pts=&Trt.; %END;
```

```
%IF %EVAL(&Ctr. > 0 AND &Trt. > 0) %THEN %DO;
```

*** The "Delta" dataset is built;

*** Each observation contains one value of the control arm and all available values of the treatment arm, at each time &time. In the "Delta" dataset differences between control and treatment values are calculated and stored;

```
Proc transpose Data=&dset. out=wilc(drop= _name_) prefix=Trt;
```

```
var &time.;
```

```
where arm = "&Treatment." and &time. ne .;
```

```
run;
```

```

Proc sql;

    create table Delta&time. as
    select a.&time. as Ctr, b.*
        from &dset. as a, wilc as b
        where a.arm = "&Control.";
    quit;

Data Delta&time.;

    set Delta&time.;
    time = "&time.";
    run;

*** The exact distribution of the Mann-Whitney statistics under null hypothesis is calculated by
resampling techniques;

%DO j=1 %TO %EVAL(&nSimul.);
%Let root=%EVAL(&seed.+&j.);
proc surveyselect data=&dset.
    method=srs n=&Ctr. out=SampleSRS noprint seed=&root.;
    where arm in ("&Control.", "&Treatment.");
run;

Proc sql noprint;

    select quote(subject)
        into :IDrandom separated by ','
        from SampleSRS;
    quit;

Data Sample;

    set &dset. (where= (arm in("&Control.", "&Treatment.")));
    if subject in(&IDrandom.) then arm = "&Control.";
    else arm = "&Treatment.";
    run;

Proc transpose Data=Sample out=MWset(drop= _name_) prefix=Trt;
    var &time.;
    where arm = "&Treatment." and &time. ne .;
    run;

```

```
Proc sql;
    create table auxMW1 as
    select a.&time. as Ctr, b.*
        from sample as a, MWset as b
        where a.arm = "&Control.";
    quit;
```

```
%LET dimens = Trt%LEFT(&Trt.);
```

```
Data auxMW2 (drop= i);
    set auxMW1;
    array rango (1:&Trt.) Trt1-&dimens.;
    MW = 0;
    do i=1 to &Trt.;
        if Ctr > rango(i) then MW = MW + 1;
    end;
    run;
```

```
Proc sql;
    create table auxMW3&j. as
        select sum(MW) as MWobs
            from auxMW2;
    quit;
```

```
Data MWaux;
    length time $ 250;
    set auxMW3&j.;
    time = "&time.";
    run;
```

```
Data MW&time.;
    set MW&time. MWaux;
    run;
```

```
%END;
```

```
Data MWexact;
    set MWexact MW&time.;
    run;
```

```

Data Delta;
    set Delta Delta&time.;
    run;
%END;
%LET count=%EVAL(&count.+1);
%END;
*** Slope differences between control and treatment arm are calculated at each time;
Data Delta (drop= i);
    set Delta;
    array Trt (1:&Pts) Trt1-Trt&Pts.;
    array diff (1:&Pts) Delta1-Delta&Pts.;
    do i=1 to &Pts.;
        if Trt(i) ne . then diff(i) = Trt(i)-Ctr;
    end;
    run;
*** Slope differences between control and treatment arm are ordered by time and value;
Data Ranking;
    set Delta;
    array diff (1:&Pts) Delta1-Delta&Pts.;
    do i=1 to &Pts.;
        Delta = diff(i);
        if Delta ne . then do; keep time delta; output Ranking; end;
    end;
    run;
Proc sort Data=Ranking;
    by time delta;
    run;
Data Ranking;
    retain rankDelta 1;
    set Ranking;
    by time;
    if first.time then rankDelta = 1;

```



```
else rankDelta = sum(rankDelta,1);
```

```
run;
```

*** The probability distribution of the Mann-Whitney statistics under null hypothesis is calculated and stored in “ditrib” dataset;

```
Proc sql;
```

```
create table ditrib as
```

```
select (count(MWobs)) as freq, (calculated freq) / &nSimul. as probMW, MWobs, time
```

```
from MWexact
```

```
group by time, MWobs
```

```
order by time, MWobs;
```

```
quit;
```

*** The cumulative distribution of the Mann-Whitney statistics under null hypothesis is calculated and stored in “rankMW” dataset;

```
Data rankMW;
```

```
retain cumProb rankMW;
```

```
set ditrib;
```

```
by time;
```

```
if first.time then do; rankMW = 1; cumProb = ProbMW; end;
```

```
else do; rankMW = sum(rankMW,1); cumProb = sum(cumProb,ProbMW); end;
```

```
run;
```

*** Useful percentiles of the Mann-Whitney statistics under null hypothesis are calculated and stored in “posMW” dataset;

```
Data posMW;
```

```
set rankMW;
```

```
by time;
```

```
if cumProb <= &alpha./2 then low = 1; else low = 0;
```

```
if 1-cumProb <= &alpha./2 then upp = 1; else upp = 0;
```

```
run;
```

*** Theorem 5.5.2.1 is applied in the following SAS statements;

```
Proc sql;
```

```
create table lowCI as
```

```
select max(MWobs) as lowCI, time
```

```

        from posMW
        where low=1
        group by time;
create table uppCI as
select min(MWobs) as uppCI, time
        from posMW
        where upp=1
        group by time;
create table CI as
        select a.time, a.lowCI, b.uppCI
                from lowCI as a left join uppCI as b
                on a.time=b.time
                order by time;
quit;

```

Proc sql;

```

create table Selection as
select a.*, b.lowCI, b.uppCI
        from Ranking as a left join CI as b
        on a.time=b.time
        order by time, rankDelta;
quit;

```

*** Credibility intervals of the additive effect are calculated by time interval;

Data CredibilityInt;

```

length typeCI $ 3;
set Selection;
if rankDelta = lowCI then do;

```

```

        typeCI = 'Low';
        CI = Delta;
        keep time typeCI Delta;
        output CredibilityInt;

```

end;

```

if rankDelta = uppCI then do;

```

```

                                typeCI = 'Upp';
                                CI = Delta;
                                keep time typeCI Delta;
                                output CredibilityInt;
                                end;

run;

Proc sort Data=CredibilityInt;
    by time typeCI;
run;

Proc transpose Data=CredibilityInt out=ListOfValues (drop= _NAME_);
    by time;
    var Delta;
    id typeCI;
run;

*** Output variables: time interval, lower estimate of the credibility interval, upper estimate of the
credible interval;

Proc sql;
    select time, Low, Upp
        from ListOfValues
        order by time;
    quit;

%MEND;

```

Appendix F

Tumor volumes (mm³) measured during the *in vivo* experiment with ML017/ET myxoid liposarcoma PDX

Arm	ID	Days from treatment start													
		0	4	7	12	15	18	21	26	29	34	36	40	43	46
Vehicle	1	342,8	403,2	693,2	1004,2	1422,6	1933,0	2244,8							
Vehicle	2	203,0	253,5	267,4	346,2	501,4	688,3	946,4							
Vehicle	3	316,5	437,8	563,1	983,6	1150,5	1467,2	1880,8							
Vehicle	4	296,9	466,3	642,9	1102,8	1403,1	2011,2	2633,9							
Vehicle	5	269,7	351,3	496,7	720,0	1074,2	1317,9	1351,3							
Vehicle	6	191,4	316,0	431,5	848,2	1412,9	2064,9	2505,8							
Vehicle	7	280,5	557,6	736,7	1256,1	1522,6	1948,0	2548,4							
Vehicle	8	215,6	320,6	468,0	779,5	1085,4	1500,8	2113,3							
Vehicle	9	163,2	274,8	346,8	577,1	779,6	943,3	1248,6							
Vehicle	10	311,7	481,4	770,6	1301,7	2193,5	2862,9	3899,9							
Vehicle	11	150,3	127,7	141,3	200,4	244,9	267,8	400,9							
Vehicle	12	211,0	287,1	622,7	967,9	838,0	604,8	564,0							
Pioglitazone	13	395,2	695,4	870,7	2060,4	2543,0	3098,3	3753,2							
Pioglitazone	14	320,9	460,4	731,1	1311,5	1734,9	1923,8	2369,9							
Pioglitazone	15	105,3	190,2	254,0	514,1	636,2	875,9	1204,4							
Pioglitazone	16	318,1	485,4	1013,9	1235,0	1505,0	1657,3	1955,5							

Arm	ID	Days from treatment start													
		0	4	7	12	15	18	21	26	29	34	36	40	43	46
Pioglitazone	17	142,2	176,7	232,1	256,7	396,0	388,4	415,5							
Pioglitazone	18	315,3	704,1	1300,9	1563,2	1712,3	1572,0	1592,3							
Pioglitazone	19	279,8	484,1	708,1	1154,4	1346,7	1378,5	1318,9							
Pioglitazone	20	184,8	354,8	547,0	768,0	890,0	935,4	1049,7							
Pioglitazone	21	162,7	255,3	504,3	683,5	914,2	991,0	1120,9							
Pioglitazone	22	254,3	393,9	716,9	997,2	1467,3	1701,1	2135,9							
Pioglitazone	23	227,6	368,5	616,8	1015,4	1089,4	1240,9	1455,0							
Trabectedin	24	119,4	131,0	193,5	198,1	232,1	165,3	211,1	252,7	313,0					
Trabectedin	25	342,4	396,6	523,7	648,5	720,8	675,5	817,7	989,4	1013,3					
Trabectedin	26	305,8	499,8	679,4	1133,5	1148,3	1367,7	2359,4	3826,2	4613,0					
Trabectedin	27	191,3	228,3	276,9	260,5	330,9	428,4	543,3	671,3	931,7					
Trabectedin	28	321,5	325,3	401,5	564,6	790,6	884,8	1259,8	1902,1	2495,9					
Trabectedin	29	275,5	483,8	610,1	771,8	887,4	949,3	1107,5	1076,9	1441,0					
Trabectedin	30	264,5	305,3	311,2	510,3	695,7	515,9	695,2	862,8	789,7					
Trabectedin	31	203,6	225,6	380,3	401,3	572,2	585,5	978,0	1679,6	2296,1					
Trabectedin	32	182,2	257,2	318,2	377,9	496,4	531,7	720,8	808,9	1186,5					
Trabectedin	33	277,5	370,4	561,8	579,5	747,6	818,4	1276,5	1428,5	1937,9					
Trabectedin	34	213,8	317,9	523,8	630,5	886,4	1151,7	1545,5	2113,4	3145,8					
Trab + Pio	35	107,7	156,8	256,8	276,4	295,2	273,7	342,0	402,2	433,9	555,6	559,0	578,1	679,8	732,9

Arm	ID	Days from treatment start													
		0	4	7	12	15	18	21	26	29	34	36	40	43	46
Trab + Pio	36	384,8	460,1	631,1	792,8	805,6	843,3	817,8	860,0	889,4	1003,4	1183,9	1420,4	1861,2	2022,9
Trab + Pio	37	330,2	522,6	702,5	1068,6	1233,8	1034,7	1209,3	1383,6	1390,7	1601,3	1567,6	1724,6	2027,8	2258,1
Trab + Pio	38	281,8	304,0	487,3	701,8	721,0	835,5	854,5	894,8	983,2	1121,8	1140,7	1367,1	1576,9	1849,5
Trab + Pio	39	251,2	306,9	339,1	344,3	369,6	348,4	379,2	291,9	346,6	434,4	423,9	449,3	560,6	616,2
Trab + Pio	40	185,6	190,7	485,7	706,9	842,3	771,6	929,5	1108,8	1207,5	1488,7	1670,4	2066,0	2246,9	2760,1
Trab + Pio	41	241,8	253,1	525,5	700,5	937,5	900,7	971,9	1170,2	1385,5	1496,1	1972,4	2553,1	2794,6	3383,7
Trab + Pio	42	133,9	143,1	264,2	309,8	366,7	438,3	471,6	538,9	544,3	558,1	697,7	690,6	904,9	953,0
Trab + Pio	43	314,7	296,8	566,8	688,5	737,1	776,8	893,1	894,4	1123,0	1244,7	1294,8	1299,2	1602,0	1473,9
Trab + Pio	44	244,3	314,3	524,6	673,0	708,7	672,6	789,9	901,5	1103,4	1291,3	1436,9	1852,9	2249,9	2185,6
Trab + Pio	45	225,2	348,7	555,8	669,2	686,0	742,4	812,1	1028,2	1168,4	1123,6	1306,4	1464,6	1699,4	2122,2