**New benchmarks to design clinical trials with advanced or metastatic liposarcoma or synovial sarcoma patients: an EORTC Soft Tissue and Bone Sarcoma Group (STBSG) meta-analysis based on a literature review for soft-tissue sarcomas**

Georgios Kantidakis, Saskia Litière, Anouk Neven, Marie Vinches, Ian Judson, Jean-Yves Blay, Eva Wardelmann, Silvia Stacchiotti, Lorenzo D’Ambrosio, Sandrine Marréaud, Winette T.A. van der Graaf, Bernd Kasper, Marta Fiocco, Hans Gelderblom

APPENDIX

Correspondence to:

Prof Hans Gelderblom,

Department of Medical Oncology,

Leiden University Medical Center, 2333 ZA,

Leiden, Netherlands

[a.j.gelderblom@lumc.nl](mailto:a.j.gelderblom@lumc.nl)

Table of Contents

[1 Details about the literature review 3](#_Toc97281434)

[2 Details about the liposarcoma (LPS) and the synovial sarcoma (SS) meta-analyses 5](#_Toc97281435)

[2.1 Study characteristics 5](#_Toc97281436)

[2.1.1 Studies for LPS meta-analyses 5](#_Toc97281437)

[2.1.2 Studies for SS meta-analyses 8](#_Toc97281438)

[2.2 Statistical considerations 12](#_Toc97281439)

[2.2.1 Unit of the analyses 12](#_Toc97281440)

[2.2.2 Meta-analysis methodology 12](#_Toc97281441)

[2.2.3 Hypothesis generating analyses 13](#_Toc97281442)

[2.2.4 Sensitivity meta-analysis 13](#_Toc97281443)

[2.2.5 Risk of bias assessment 13](#_Toc97281444)

[2.2.6 Proportions and transformation function 14](#_Toc97281445)

[2.3 Statistical inference for liposarcoma (LPS) 14](#_Toc97281446)

[2.3.1 Analysis for first-line patients at 3 months 14](#_Toc97281447)

[2.3.2 Analysis for first-line patients at 6 months 18](#_Toc97281448)

[2.3.3 Analysis for pre-treated patients at 3 months 21](#_Toc97281449)

[2.3.4 Analysis for pre-treated patients at 6 months 24](#_Toc97281450)

[2.4 Statistical inference for synovial sarcoma (SS) 28](#_Toc97281451)

[2.4.1 Analysis for first-line patients at 3 months 28](#_Toc97281452)

[2.4.2 Analysis for first-line patients at 6 months 31](#_Toc97281453)

[2.4.3 Analysis for pre-treated patients at 3 months 33](#_Toc97281454)

[2.4.4 Analysis for pre-treated patients at 6 months 35](#_Toc97281455)

[3 References 39](#_Toc97281456)

# Details about the literature review

**Subject index terms**:

Soft-tissue sarcoma, phase II (2), phase III (3), phase IV (4), metastatic, advanced, clinical trial(s).

**Type of studies to be included/excluded**:

Eligible study designs included randomized controlled or non-randomized clinical trials of phase II or III as well as prospective real-life studies.

Case control studies, case series, review papers, clinical trials of phase I, reports, pooled analyses, sub-studies, and ecological studies were excluded.

**Condition or domain being studied**:

Systemic therapy (chemotherapy, targeted, immunotherapy or cell therapy) in advanced non-resectable or metastatic soft tissue sarcoma (STS) for first or later lines of treatment.

**Participants/Population**:

Population of the trials were adults with a diagnosis of advanced or metastatic STS, non-operable who underwent systemic therapy (reported).

**Electronic search strategy (according to PRISMA guidelines** [1]**)**:

The search was restricted to publications in the period: 1 January 2003 – 31 December 2018 using the English language.

**Electronic database search**: MEDLINE through PubMed database.

The search was performed combining two separate searches to be as general as possible. The MeSH term sarcoma was preferred to capture all sarcoma subtypes and based on the screening process papers regarding soft-tissue sarcomas were selected. The algorithms are presented below:

("sarcoma"[MeSH Terms] OR "sarcoma"[All Fields]) AND advanced [All Fields] AND (Clinical Trial[ptyp] AND ("2003/01/01"[PDAT] : "2018/12/31"[PDAT]) AND "humans"[MeSH Terms] AND English[lang]) 🡪 419 potential articles were identified.

("sarcoma"[MeSH Terms] OR "sarcoma"[All Fields]) AND ("secondary"[Subheading] OR "secondary"[All Fields] OR "metastatic"[All Fields]) AND (Clinical Trial[ptyp] AND ("2003/01/01"[PDAT] : "2018/12/31"[PDAT]) AND "humans"[MeSH Terms] AND English[lang]) 🡪 512 potential articles were identified.

Combining both searches, the total number of potentially eligible papers was **745** after removal of 186 duplicate results between the two searches.

**First screening step (title and abstract – without details)**

**Inclusion criteria**:

* Diagnosis of a primary STS.
* Advanced or metastatic non-resectable disease.
* Tumour characteristics:

Any STS histology including but not limited to angiosarcoma, leiomyosarcoma, liposarcoma, synovial sarcoma, rhabdomyosarcoma, epithelioid sarcoma. Also, any anatomical site was eligible.

* Any systemic treatment (chemotherapy, target, immunotherapy or cell therapy) of advanced or metastatic STS in first-line or pre-treated setting.
* Publications in the period 1 January 2003 – 31 December 2018.
* The language of the published paper was English.

**Exclusion criteria**:

* Paediatric population.
* Surgery or radiotherapy for localized disease, limb perfusion, electro chemotherapy, supportive care interventions.
* Articles dedicated exclusively to GIST, bone sarcomas or other types of cancer. Note that papers that study STS **and** other cancers were included (for example STS and bone sarcomas).
* Early phase trials (Phase I, Phase I-II).
* Diagnostic or biomarker trials.
* No clinical data/trial (e.g. review papers, position papers without clinical data).
* Retrospective clinical data.
* Sub-analysis or pooled analysis papers.

For the last criterion, it was examined if the main study (in case of a subgroup analysis) or the pooled studies (in case of a pooled analysis) were included in the search. If any study was missing, it was added manually.

**Results of step 1 (reported according to PRISMA guidelines** [1]**):**

According to this screening step, 165 / 745 papers were deemed potentially eligible. In addition, one extra paper was added manually by ad-hoc checking of sub-studies, pooled analyses (themselves excluded). Thus, **166** potentially eligible articles were examined during the second screening step. Subsequently, to perform the second screening step the full-text of the articles was retrieved from PubMed (by MV and GK) or by EORTC HQ communication group (in cases where the full-text was not accessible) - buying the papers from the journals published.

**Screening step 2 (abstract – with details, paper – without details)**

**Inclusion criteria**:

The criteria of inclusion were the same with the first screening phase.

**Exclusion criteria**:

The additional exclusion criterion here was:

Papers of studies that

* Include other types of cancers than STS **and** in which STS were not analysed separately (pooled analysis for all types of cancer).

**Results of step 2:**

Seven papers were removed based on the additional exclusion criterion leading to 159 clinical trials in total.

# Details about the liposarcoma (LPS) and the synovial sarcoma (SS) meta-analyses

* 1. **Study characteristics**

### **Studies for LPS meta-analyses**

In this subsection we provide descriptive statistics regarding the **25 trials** with LPS (PFS estimates at 3-6 months could not be retrieved for 10/35 trials) included in the meta-analysis, and the corresponding **34 rows** (one row per treatment arm and line) of the long-transformed version of the database. For instance, the phase II trial of Maurel et al. (2009) [2] included 2 treatment arms either doxorubicin monotherapy or doxorubicin in combination with Ifosfamide corresponding to 2 rows of the excel database (both regarding first-line treatment). Note that the Bui-Nguyen study [3] contained 3 treatment arms: Doxorubicin, Trabectedin 3h and Trabectedin 24h. The 3 hours regimen has been excluded from meta-analyses because of the limited number of LPS patients (n = 6). The Gelderblom study [4] contained 2 treatment arms: Doxorubicin and Brostallicin. The Doxorubicin arm was excluded from the analysis because it did not reach the predetermined number of patients (n = 9). The Kawai study [5] included a best supportive care arm which was excluded from the meta-analyses.

In the table below, year of study activation (year that first patient was recruited) is presented for the 25 trials in the STBSG database. Most common years of activation at trial level were 2011 and 2012 (24% and 16%).

|  |  |  |
| --- | --- | --- |
| **Year of study activation** | **N per trial (%)** | **N per row (%)** |
| 2001 | 1 (4%) | 1 (2.94%) |
| 2003 | 2 (8%) | 4 (11.76%) |
| 2004 | 1 (4%) | 1 (2.94%) |
| 2005 | 1 (4%) | 1 (2.94%) |
| 2006 | 1 (4%) | 1 (2.94%) |
| 2007 | 2 (8%) | 2 (5.88%) |
| 2008 | 2 (8%) | 3 (8.82%) |
| 2010 | 3 (12%) | 4 (11.76%) |
| 2011 | 6 (24%) | 11 (32.35%) |
| 2012 | 4 (16%) | 4 (11.76%) |
| 2013 | 1 (4%) | 1 (2.94%) |
| 2015 | 1 (4%) | 1 (2.94%) |
| Total | 25 (100%) | 34 (100%) |

|  |  |  |
| --- | --- | --- |
| **Year of study publication** | **N trial (%)** | **N row (%)** |
| 2009 | 3 (12%) | 4 (11.76%) |
| 2011 | 2 (8%) | 2 (5.88%) |
| 2013 | 3 (12%) | 3 (8.82%) |
| 2014 | 3 (12%) | 5 (14.71%) |
| 2015 | 3 (12%) | 4 (11.76%) |
| 2016 | 5 (20%) | 8 (23.53%) |
| 2017 | 6 (24%) | 8 (23.53%) |
| Total | 25 (100%) | 34 (100%) |

|  |  |  |
| --- | --- | --- |
| **Phase** | **N per trial (%)** | **N per row (%)** |
| 2 | 17 (68%) | 19 (55.88%) |
| 2 | 3 | 1 (4%) | 2 (5.88%) |
| 3 | 6 (24%) | 12 (35.29%) |
| 4 | 1 (4%) | 1 (2.94%) |
| Total | 25 (100%) | 34 (100%) |

17 trials were phase 2 (68%) and 8 (32%) greater than 2. From the 34 lines in long format (one line per treatment combination), 19 were of phase 2 (55.88%) and 15 of greater than 2 (44.12%).

|  |  |  |
| --- | --- | --- |
| **Study design** | **N per trial (%)** | **N per row (%)** |
| randomized | 11 (44%) | 19 (55.88%) |
| non-randomized | 14 (56%) | 15 (44.12%) |
| Total | 25 (100%) | 34 (100%) |

From the 34 excel lines analyzed, 19 were randomized (55.88%) and 15 non-randomized (44.12%).

|  |  |  |
| --- | --- | --- |
| **Pathological review** | **N per trial (%)** | **N per row (%)** |
| local | 9 (36%) | 12 (35.29%) |
| central | 10 (40%) | 13 (38.24%) |
| unclear | 6 (24%) | 9 (26.47%) |
| Total | 25 (100%) | 34 (100%) |

Pathological review was unclear (whether local or central) for 6 trials.

|  |  |  |
| --- | --- | --- |
| **Primary endpoint** | **N per trial (%)** | **N per row (%)** |
| Progression-Free Rate/Survival Rate at 3 months | 9 (36%) | 10 (29.41%) |
| Progression-Free Survival | 7 (28%) | 10 (29.41%) |
| Overall Survival | 4 (16%) | 8 (23.53%) |
| Clinical Benefit Rate | 3 (12%) | 3 (8.82%) |
| Progression-Free Rate/Survival Rate at 6 months | 1 (4%) | 2 (5.88%) |
| Response Rate | 1 (4%) | 1 (2.94%) |
| Total | 25 (100%) | 34 (100%) |

The most common primary endpoints per trial were Progression-Free Rate/Survival Rate at 3 months and Progression-Free Survival (9 and 7 times, respectively).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Summary statistics for % of evaluable LPS patients** | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| Over total patients registered | 4.28 | 9.22 | 18.42 | 25.90 | 30.77 | 100.00 |
| Over LPS patients registered | 90.32 | 98.33 | 100.00 | 98.46 | 100.00 | 100.00 |

Liposarcomas (evaluable patients only) represent on average 25.90% of the total sample size for soft-tissue sarcomas in these trials – a non-negligible percentage. Four of the 25 trials (Dickson 2013 and 2016, Toulmonde 2015 and Samuels 2017) were designed exclusively for LPS [6–9].

Followingly, we show a table regarding the 34 treatments considered per row for this review.

|  |  |  |
| --- | --- | --- |
| **Drugs or drug combinations used** | **N (%)** | **Recommended according to ESMO 2021 guidelines** [10]**?** |
| Doxorubicin | 5 (14.71%) | Yes (for first-line) |
| Trabectedin | 4 (11.76%) | No (for first-line)  Yes (for pre-treated) |
| Doxorubicin + Ifosfamide | 3 (8.82%) | Yes (for first-line) |
| Eribulin | 3 (8.82%) | Yes (for pre-treated) |
| Palbociclib | 3 (8.82%) | No (for first-line)  No (for pre-treated) |
| Dacarbazine | 2 (5.88%) | No (for pre-treated) |
| Pazopanib | 2 (5.88%) | No (for pre-treated) |
| Aplidin | 1 (2.94%) | No (for pre-treated) |
| Brostallicin | 1 (2.94%) | No (for first-line) |
| Cixutumumab | 1 (2.94%) | No (for pre-treated) |
| Dasatinib | 1 (2.94%) | No (for pre-treated) |
| Docetaxel + Gemcitabine | 1 (2.94%) | No (for first-line) |
| Doxorubicin + Evofosfamide | 1 (2.94%) | No (for first-line) |
| Imatinib | 1 (2.94%) | No (for pre-treated) |
| Panobinostat | 1 (2.94%) | No (for pre-treated) |
| Pembrolizumab | 1 (2.94%) | No (for pre-treated) |
| Regorafenib | 1 (2.94%) | No (for pre-treated) |
| Ridaforolimus | 1 (2.94%) | No (for pre-treated) |
| Trabectedin 24h | 1 (2.94%) | No (for first-line) |
| Total | 34 (100%) | //////////////////////////// |

The most common drugs on these trials were Doxorubicin, either alone or in combination, and Trabectedin (5 times).

|  |  |  |
| --- | --- | --- |
| **Treatment line** | **N per trial (%)** | **N per row (%)** |
| first | 8 (32%) | 14 (41.18%) |
| pre-treated | 17 (68%) | 20 (58.82%) |
| Total | 25 (100%) | 34 (100%) |

20 excel lines were based on pre-treated population (from 17 studies) and 14 excel lines for first-line (8 clinical trials).

|  |  |
| --- | --- |
| **Regimen recommended?** | **N (%)** |
| No | 20 (58.82%) |
| Yes | 14 (41.18%) |
| Total | 34 (100%) |

From the 34 treatment lines selected (25 trials), 20 were non-recommended (58.82%) and 14 were recommended (41.18%).

### **Studies for SS meta-analyses**

In this subsection we provide descriptive statistics regarding the **13 trials** with SS (PFS estimates at 3-6 months could not be acquired for 3/13 trials) included in the meta-analysis, and the corresponding **15 rows** (one line per treatment arm) of the long-transformed version of the database. The Blay (2014) study [11] included a doxorubicin + ifosfamide arm which was removed from the database as it did not reach the predetermined sample size (n = 9 <10).

In the table below, year of study activation (year that first patient was recruited) is presented for the 13 trials in the STBSG database. Most common year of activation at trial level was 2008 (30.77%).

|  |  |  |
| --- | --- | --- |
| **Year of study activation** | **N per trial (%)** | **N per row (%)** |
| 2001 | 1 (7.69%) | 1 (6.67%) |
| 2002 | 1 (7.69%) | 1 (6.67%) |
| 2003 | 1 (7.69%) | 2 (13.33%) |
| 2005 | 1 (7.69%) | 1 (6.67%) |
| 2007 | 1 (7.69%) | 1 (6.67%) |
| 2008 | 4 (30.77%) | 4 (26.67%) |
| 2011 | 1 (7.69%) | 2 (13.33%) |
| 2012 | 1 (7.69%) | 1 (6.67%) |
| 2013 | 1 (7.69%) | 1 (6.67%) |
| 2015 | 1 (7.69%) | 1 (6.67%) |
| Total | 13 (100%) | 15 (100%) |

Most common year of publication at trial level was 2017 (23.08%).

|  |  |  |
| --- | --- | --- |
| **Year of study publication** | **N trial (%)** | **N row (%)** |
| 2008 | 1 (7.69%) | 1 (6.67%) |
| 2009 | 2 (15.38%) | 2 (13.33%) |
| 2011 | 1 (7.69%) | 1 (6.67%) |
| 2012 | 1 (7.69%) | 1 (6.67%) |
| 2013 | 1 (7.69%) | 1 (6.67%) |
| 2014 | 2 (15.38%) | 3 (20%) |
| 2015 | 1 (7.69%) | 1 (6.67%) |
| 2016 | 1 (7.69%) | 1 (6.67%) |
| 2017 | 3 (23.08%) | 4 (26.67%) |
| Total | 13 (100%) | 15 (100%) |

|  |  |  |
| --- | --- | --- |
| **Phase** | **N per trial (%)** | **N per row (%)** |
| 2 | 8 (61.54%) | 8 (53.33%) |
| 2|3 | 0 (0%) | 0 (0%) |
| 3 | 4 (30.77%) | 6 (40%) |
| 4 | 1 (7.69%) | 1 (6.67%) |
| Total | 13 (100%) | 15 (100%) |

8 trials were phase 2 (61.54%), 4 (30.77%) phase 3 and 1 phase 4 (7.69%). From the 15 lines in long format (one line per treatment combination), 8 were of phase 2 (53.33%) and 7 of greater than 2 (46.67%).

|  |  |  |
| --- | --- | --- |
| **Study design** | **N per trial (%)** | **N per row (%)** |
| randomized | 5 (38.46%) | 7 (46.67%) |
| non-randomized | 8 (61.54%) | 8 (53.33%) |
| Total | 13 (100%) | 15 (100%) |

From the 15 excel lines analyzed, 7 came from randomized and 8 from non-randomized studies.

|  |  |  |
| --- | --- | --- |
| **Pathological review** | **N per trial (%)** | **N per row (%)** |
| local | 6 (46.15%) | 7 (46.67%) |
| central | 4 (30.77%) | 5 (33.33%) |
| unclear | 3 (23.08%) | 3 (20%) |
| Total | 13 (100%) | 15 (100%) |

Pathological review was locally performed for 6 trials, centrally for 4 trials and was unclear for 3 trials.

|  |  |  |
| --- | --- | --- |
| **Primary endpoint** | **N per trial (%)** | **N per row (%)** |
| Progression-Free Rate/Survival Rate at 3 months | 4 (30.77%) | 4 (26.67%) |
| Progression-Free Survival | 4 (30.77%) | 4 (26.67%) |
| Overall Survival | 2 (15.38%) | 4 (26.67%) |
| Response Rate | 2 (15.38%) | 2 (13.33%) |
| Clinical Benefit Rate | 1 (6.67%) | 1 (6.67%) |
| Total | 13 (100%) | 15 (100%) |

The most common primary endpoints per trial were PFSR at 3 months and PFS.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Summary statistics for % of evaluable SS patients** | **Min.** | **1st Qu.** | **Median** | **Mean** | **3rd Qu.** | **Max.** |
| Over total patients registered | 2.66 | 7.80 | 11.63 | 21.33 | 21.09 | 95.83 |
| Over SS patients registered | 95.45 | 100.00 | 100.00 | 99.07 | 100.00 | 100.00 |

Synovial sarcoma (evaluable patients only) represent on average 21.33% of the total sample size in these trials – a non-negligible percentage. The trial by Ray-Coquard was designed exclusively for synovial sarcoma [12].

Followingly, we show a table regarding the 15 treatment rows for SS.

|  |  |  |
| --- | --- | --- |
| **Drugs or drug combinations used** | **N (%)** | **Recommended according to ESMO 2021 guidelines** [10]**?** |
| Doxorubicin | 2 (13.33%) | Yes (for first-line) |
| Pazopanib | 2 (13.33%) | Yes (for pre-treated) |
| Trabectedin | 2 (13.33%) | No (for first-line)  Yes (for pre-treated) |
| Cixutumumab | 1 (6.67%) | No (for pre-treated) |
| Cyclosphosphamide + Fludarabine + TCR transduced cells | 1 (6.67%) | No (for pre-treated) |
| Doxorubicin + Evofosfamide | 1 (6.67%) | No (for first-line) |
| Doxorubicin + Ifosfamide | 1 (6.67%) | Yes (for first-line) |
| Eribulin | 1 (6.67%) | No (for pre-treated) |
| Gefitinib | 1 (6.67%) | No (for pre-treated) |
| Imatinib | 1 (6.67%) | No (for pre-treated) |
| Pembrolizumab | 1 (6.67%) | No (for pre-treated) |
| Regorafenib | 1 (6.67%) | No (for pre-treated) |
| Total | 15 (100%) | //////////////////////////// |

The most common drug / drug combinations were Doxorubicin, Pazopanib and Trabectedin.

|  |  |  |
| --- | --- | --- |
| **Treatment line** | **N per trial (%)** | **N per row (%)** |
| first | 3 (23.08%) | 5 (33.33%) |
| pre-treated | 10 (76.92%) | 10 (66.67%) |
| Total | 13 (100%) | 15 (100%) |

5 excel lines were based on first-line treatment (from 3 studies) and 10 excel lines were based on pre-treated population (10 clinical trials). Mixed treatment lines were not included.

|  |  |
| --- | --- |
| **Regimen recommended?** | **N (%)** |
| No | 9 (60%) |
| Yes | 6 (40%) |
| Total | 15 (100%) |

From the 15 treatment lines selected (13 trials), 6 were recommended (40%).

* 1. **Statistical considerations**

### **Unit of the analyses**

The unit for all analyses was the estimated **proportion** (number of cases / sample size) for each drug regimen. Equivalently, the estimated **rate** is the proportion\*100.

### **Meta-analysis methodology**

It is necessary to access for each study not only the effect size (in terms of PFS at 3/6 months) but also the sample size of LPS / SS. To elaborate on this, for RR (as an example) the CI can be easily calculated using the sample size and assuming a binomial 95% CI based on the observed cases for a regimen. However, for the PFS as the variance of the percentages (calculated under Kaplan-Meier or binomial distribution) is rarely provided in the papers, we estimated the number of cases (pseudocases) for the LPS / SS sample size and we used them to approximate the variance with a 95% binomial distribution at 3/6 months ignoring censoring. A PFS proportion is directly the number of (pseudo)cases at 3/6 moths divided by the sample size (not the Kaplan-Meier estimate). For some studies, we used the PFS at 3/6 months estimated by KM to find the equivalent number of pseudocases. For example, assuming that PFS at 3 months is 0.33 with KM methodology for 34 LPS patients, the estimated number of pseudocases is 11 (the closest integer with a value of almost 0.33). A 95% binomial CI is actually more conservative than a 95% confidence interval provided with the Kaplan-Meier estimator (the latter is narrower as it takes into account the censoring mechanism).

To quantitatively synthesize the findings of the 25 trials (7 1L, 17 2L+, 1 both) for LPS and the 13 trials (3 1L and 10 2L+) for SS, we employed a random-effects model per row; suggested by DerSimonian and Laird for meta-analysis of clinical trials [13,14]. This model assumes that the true effect of each study (here each row – drug or drug combination per treatment line - represents a study) is normally distributed. A second assumption made is that effect sizes of drugs or drug combinations are independent, even if they come from the same study. Here the majority of the trials were single arm (14/25 for LPS, 8/13 for SS). For randomized studies, there might be some dependence as treatment arms were designed for exactly the same population (also patients were recruited in the same centers). Outcome variable is the summary proportion; the weighted average of the observed effect sizes of individual lines. Weighting is calculated using the inverse of the total variance (within + between). The main idea of the model is that studies will differ in the mixes of participants and in the implementations of the interventions – among other reasons. Therefore, a different effect size will be underlying each study [27]. Variation in each study can be quantified through the true variation in the effect size and the sampling error. Effect sizes (yi) and the sampling variance (vi) are calculated per row. True effect sizes are not fixed but normally distributed (both within and between study variance is taken into account). It is useful to mention here that a random-effects model provides generalizability of the findings for new trials, and that subgroup meta-analysis is actually performing a meta-regression for a dichotomous predictor.

The drugs used in the studies were split into two groups (recommended against non-recommended); using as criterion the ESMO 2021 guidelines [10] (variable “Is experimental drug standard of care for LPS / SS?”). For both groups, the corresponding weight of the effect size of each intervention was estimated using the inverse of the total variance (summary of between and within study variance). In addition, a mean effect size was estimated for each group together with the 95% confidence interval (CI). Larger studies were given more weight and thus their effect size had greater impact on the overall mean effect.

Moreover, we performed moderator analyses in an attempt to explain between study variability. Subgroup analyses are conducted with a mixed effects model. The random component of the model is used to combine the study effects (within each subgroup) and the fixed component of the model is used to test whether these effects vary significantly across the subgroups. Categorical moderators are study characteristics which can account for between-study heterogeneity. We evaluated different moderators: dichotomized time period of study activation, phase of the trial, study design and a dichotomized data-driven sample size separately in univariate models and we also tested significant moderators in a multivariate model including drug regimens to see if they can explain part of the variation. Residual heterogeneity is reported (remaining variability between the studies not accounted for by the moderators).

### **Hypothesis generating analyses**

* To compare different subgroups of patients (drug regimen recommended or not). To do so, we performed comparisons directly on the differences in means for the endpoint assuming that there is no difference between the groups. We computed a Z-statistic (a Wald type statistic) based on the mean differences. This test determined whether there is a statistically significant benefit for the recommended group of drugs. A small p-value of this test demonstrates a statistically significant difference between the groups. Of note is that this is equivalent to the omnibus chi-square test for moderators when a single moderator is tested.
* To investigate variation of the effect size between the studies of the subgroups. Generally, the observed variation in the studies is attributed to the true variation and to random error. We estimated the -statistic that is insensitive to the number of studies [15]. The -statistic is a signal-to-noise ratio. It measures the true heterogeneity of the observed effects (true between studies dispersion) compared to the total variation (total dispersion). As this statistic moves away from 0% towards 100%, it can be assumed that some of the variance is real.

### **Sensitivity meta-analysis**

A sensitivity analysis for outliers and influential studies is performed. In a meta-analysis, there is a small number of studies that can influence excessively the summary effect in an undesirable way (not demonstrating the general trend).

Possible outlier studies were detected with inspection of the forest plot of studies (studies where the effect estimate and its confidence interval do not overlap with the overall group estimate or its confidence interval). We also utilized a series of diagnostics (Baujat plot, influence diagnostics, GOSH plots, normal QQ-plots)[16] that have been developed over the years and can substantially contribute in identifying influential studies. Here, we present only the Baujat plot [17]. This diagnostic plot is used to detect sources of heterogeneity in meta-analysis. On the x-axis the reader can see the contribution of each study to the overall Q-test statistic for heterogeneity versus the influence of each study on the summary effect on the y-axis.

As part of this sensitivity analysis, potentially influential studies *were removed* to check how large their impact was for the pooled estimates and the group difference. Small differences in results demonstrate findings which are robust to outliers and influential studies.

### **Risk of bias assessment**

Meta-analyses are susceptible to *publication bias.* In some cases, studies that report relatively high effect sizes are more likely to be published than studies which report lower effect sizes. This bias in the literature (even if it is less frequent for clinical trials) can affect the conclusions of the meta-analyses. We examined bias using methods which are common for meta-analyses.

We used funnel plots to investigate risk of publication bias through assessment of the association between study effect and study size [18,19]. X-axis provides the residual values whereas Y-axis a study size measure (standard error or sample size). In general, publication bias translates to studies with null or undersized effects not being present. Funnel plots detect asymmetry (e.g. when there are many more circles at the bottom). However, asymmetry does not directly imply publication bias as it is also associated with citation bias and heterogeneity in the true effects. We show a contour enhanced funnel plot. This graph is centered at residual value 0.5 (no effect under the null hypothesis) with shaped regions distinguishing various levels of statistical significance.

As part of formal tests, we used regression tests. Those are *the rank correlation test* for funnel plot asymmetry and *the Egger’s regression test* for funnel plot asymmetry [20,21]. The first examines whether the observed outcomes and the corresponding sampling variances are correlated using linear regression. High correlation implies funnel plot asymmetry which could suggest publication bias. The second tests the relationship between the observed outcomes and the chosen predictor. Again such relationship implies asymmetry in the funnel plot which could imply publication bias [22]. Egger’s regression test is better for smaller number of studies.

### **Proportions and transformation function**

In case that the observed proportions get further from 0.5 and approach closer to the margins - being smaller than 0.2 or larger than 0.8 – their distribution is increasingly skewed. In such scenario a transformation of the raw rates is necessary. Appropriate transformation methods include the logit and the Freeman-Tukey double arcsine transformation [23,24]. For this project, we applied the logit (log-odds) transformation which is very popular for meta-analyses. The double arcsine transformation would be more appropriate if extreme proportions were observed (for example in a meta-analysis of RR). Back-transformations of the summary effect were also performed to obtain the effects on the actual scale.

* 1. **Statistical inference for liposarcoma (LPS)**

Here, we present at separate subsections the risk of bias assessment, corresponding meta-analyses, diagnostics for influential studies, sensitivity meta-analysis, and meta-regressions for the studies available.

### **Analysis for first-line patients at 3 months**

Our database consists of 8 trials on first-line treatment, corresponding to 14 different therapeutic combinations (14 lines in long format).

**Risk of bias assessment**

In this subsection, we investigate publication bias based on a funnel plot and 2 formal tests. We show the contour enhanced funnel plot in fig 1. Graph is centered at residual value 0.5 (no effect under the null hypothesis). There are shaped regions for various levels of statistical significance for the points (see the label). Studies are equally spread in the graph but can be found mostly in the upper part (lower standard error). Two studies are in the grey zone (0.01 ≤ p ≤ 0.05) one on the left side and one on the right side. There might be some publication bias here.

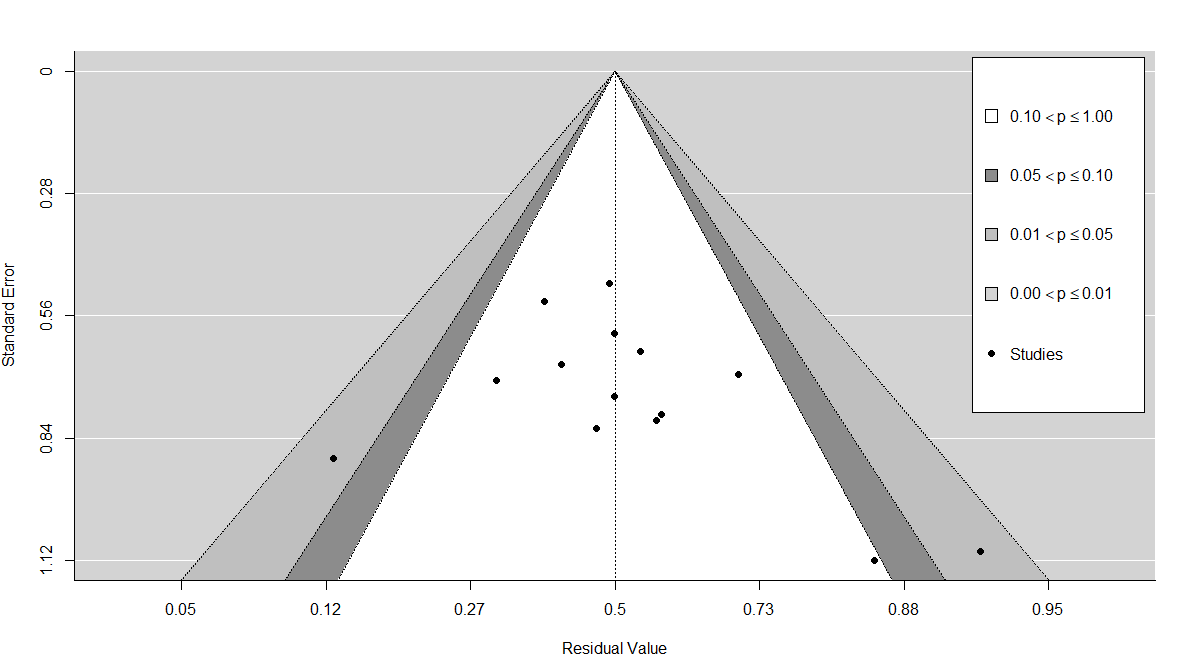


Fig. A1. Contour enhanced funnel plot of LPS for PFS at 3 months (first-line patients).

Moreover, we performed 2 formal tests: (1) the rank correlation test for funnel plot asymmetry, and (2) Egger’s regression test for funnel plot asymmetry. Here, with merely 14 lines in our database, both tests should be interpreted with caution. Rank correlation test provided a p-value of 0.15 and Egger’s regression test also a p-value of 0.13. There is no indication of publication bias from the formal tests.

**Diagnostics for influential studies**

We begin with the examination of the externally standardized residuals of the studies. Z values > 2 or 3 show an influential outlier (its exclusion may lead to significant changes in the fitted meta-analytic model). Study lines 12 (Blay 2014: Trabectedin [11]) and 14 (Gelderblom 2014:Brostallicin [4]) had the highest z value of 2.41 and -2.32, respectively (these studies could be influential).

We used the Baujat plot (figure 3) as a diagnostic tool to detect sources of heterogeneity and potentially influential studies. Study with number 12 (Blay 2014: Trabectedin) is projected on the right corner (contributes the most to heterogeneity and is the second influential after line 14).

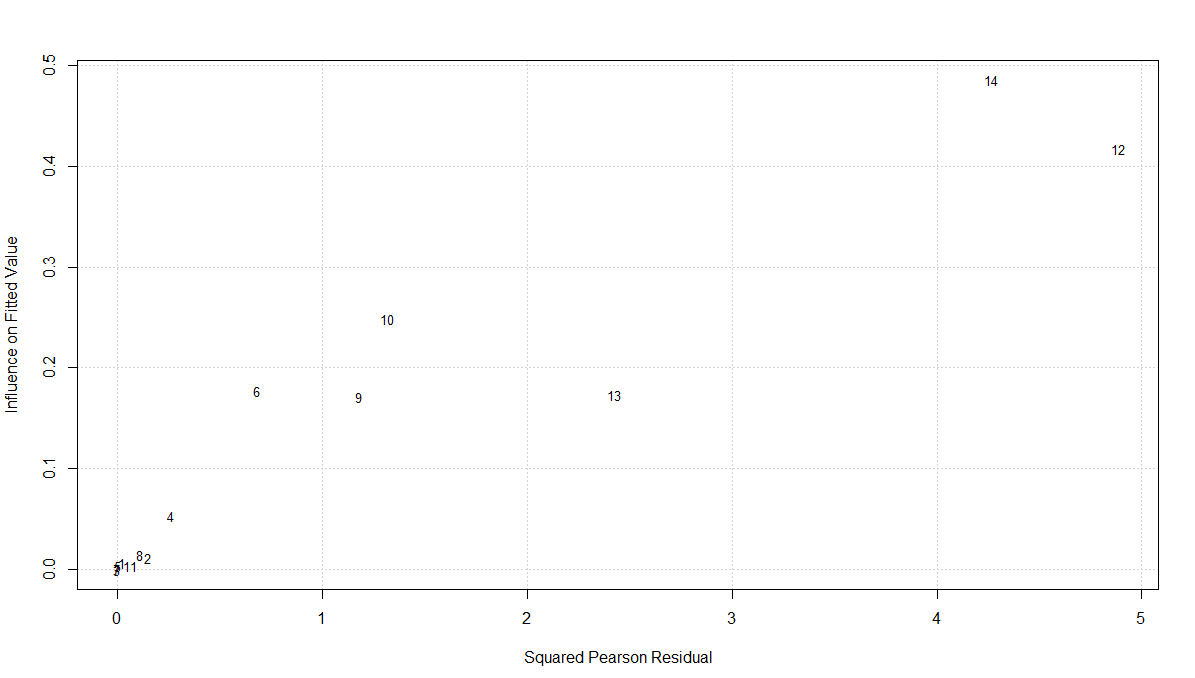


Fig. A2. Baujat plot of the LPS studies for PFS at 3 months (first-line patients).

**Sensitivity meta-analysis**

To investigate the effect of ‘Blay 2014: Trabectedin’ which was detected in previous section, we performed a sensitivity analysis removing this treatment arm from the analysis and re-fitting the random-effects model.

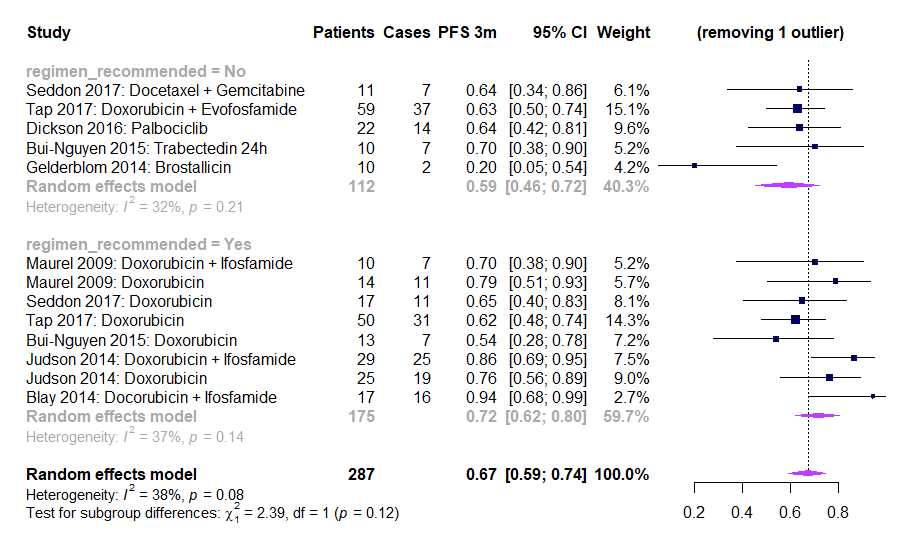


Fig. A3. Forest plot of LPS for PFS at 3 months excluding ‘Blay 2014: Trabectedin' (first-line patients).

There is a decrease in the overall effect size of the meta-analysis from 0.69 (0.60 – 0.77) to 0.67 (0.59 – 0.74). Overall heterogeneity dropped from 48% (p=0.02) to 38% (p = 0.08). This treatment arm was removed from the non-recommended treatment regimens. The estimate of the subgroup has been recalculated to 0.59 (0.46 – 0.72) from 0.64 (0.48 – 0.77). There is a substantial decrease in the group heterogeneity to 32% (from 60%). Findings indicate that ‘Blay 2014: Trabectedin' might be influential study. Again, the difference between a regimen recommended versus a non-recommended is not statistically significant based on moderator’s test (p = 0.12). Note that Trabectedin might not be recommended for first-line LPS (of any kind) but has demonstrated high activity in myxoid LPS patients (which was the LPS subtype in this trial).

**Meta-regressions**

We also examined the effect of potential moderators on the effect size through meta-regressions. First, we evaluate them separately in univariate models. If any moderators are found significant, we then test them simultaneously in a single model including the regimen recommended (variable of main interest).

For these meta-regressions we use the moderators presented below. For year of activation and sample size, we performed the dichotomization based on the median values of the entire database for LPS. We only perform analyses for moderators including more than two studies.

|  |  |
| --- | --- |
| **Moderator** | **Dichotomization** (first group reference, count of study rows in parenthesis) |
| Phase of the trial | 2 (4), greater than 2 (10) |
| Study design | Randomized (13), non-randomized (1) |
| Year of activation | 2000-2010 (9), 2011-2015 (5) |
| Sample size | 10-24 (10), >24 (4) |

Here there is only one non-randomized study so we did not consider this moderator. In the table below, we report the effect of moderators (estimate) and the residual heterogeneity in the random-effects model. We also show the p-value of the omnibus test for the effect of the moderator as well as the test for residual heterogeneity.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Moderator**  **(univariate analysis)** | **Estimated PFS (vs ref level)** | **Test for moderator**  **P-value** | **Residual heterogeneity (I2)** | **Test for residual heterogeneity**  **P-value** |
| Phase of the trial:  greater than 2 | 0.72 (vs 0.61) | 0.30 | 50.42% | 0.02 |
| Year of activation:  2011-2015 | 0.62 (vs 0.74) | 0.14 | 42.05% | 0.05 |
| Sample size: >24 | 0.71 (vs 0.68) | 0.73 | 51.72% | 0.02 |

For the phase of the trial, residual heterogeneity is moderate/high (I2 = 50.42%, p = 0.02). Studies of phase greater than 2 increase the effect size. The omnibus test for the effect of the moderator based on a chi-squared distribution with 1 degree of freedom provided a p-value of 0.30, which is not significant.

For the year of activation, residual heterogeneity is moderate (I2 = 42.05%, p = 0.05). A year of activation between 2011 and 2015 has a negative effect on the effect size, but not significant against 2000-2010 (p = 0.14).

For sample size, residual heterogeneity is again moderate/high (I2 = 51.72%, p = 0.02). A sample size above 24 patients has a positive effect on the observed proportion versus ≤ 24 which is not significant (p = 0.73).

As there were not significant moderators in the univariate models, we did not perform a multivariate meta-regression to adjust regimen recommended with phase, year of activation or sample size.

### **Analysis for first-line patients at 6 months**

Our database consists of 8 trials covering 14 different therapeutic combinations.

**Risk of bias assessment**

Again, we investigate publication bias with a funnel plot and 2 formal tests. The contour enhanced funnel plot in fig. 5 detected one study in the light grey zone of the right side (0.00 ≤ p ≤ 0.01) and another study on the dark grey zone of the right side (0.05 ≤ p ≤ 0.10). There is a chance for some publication bias.

Next, we examine the risk of bias more formally. The p-values from the rank correlation and the Egger’s regression tests were 0.67 and 0.13, respectively. There is no indication of publication bias. Note that these results could have been affected by small number of studies (14 rows for PFS at 6 months and first-line patients).

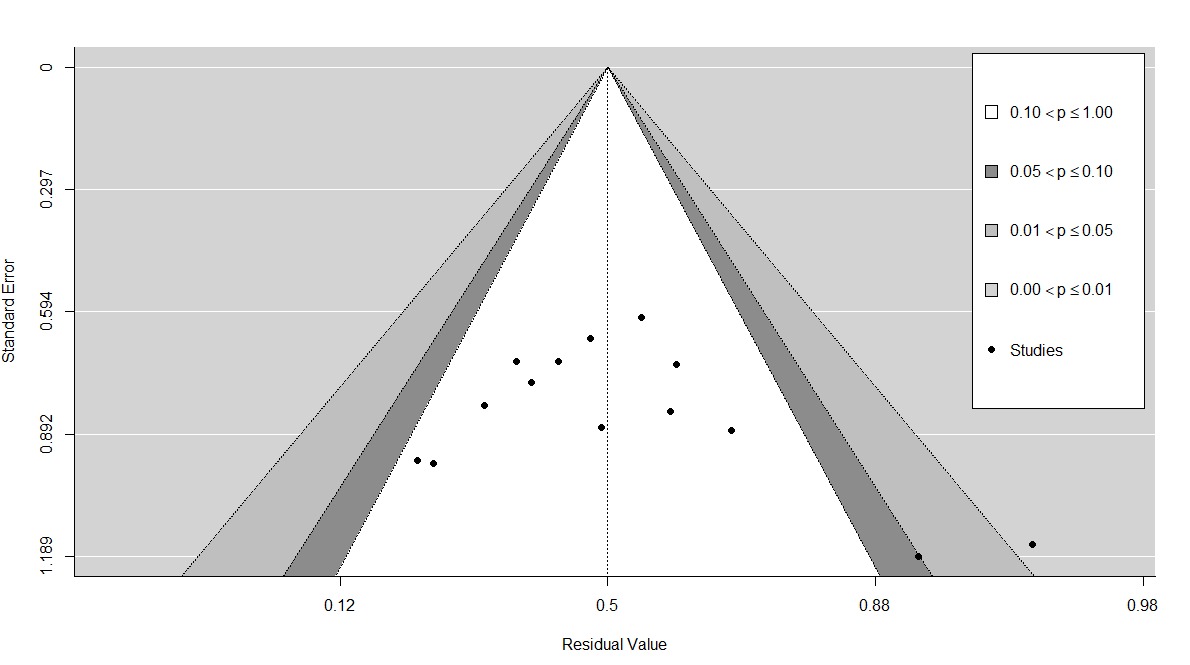


Fig. A4. Contour enhanced funnel plot of LPS for PFS at 6 months (first-line patients).

**Diagnostics for influential studies**

Externally standardized residuals produced the highest z-value of 2.90 for the study of ‘Blay 2014: Trabectedin’ [11]. This suggests that it might be an influential outlier.

The Baujat plot in fig. 7 projects line with number 12 (Blay 2014: Trabectedin) on the right corner. This study contributes to heterogeneity and is the most influential for the meta-analysis.

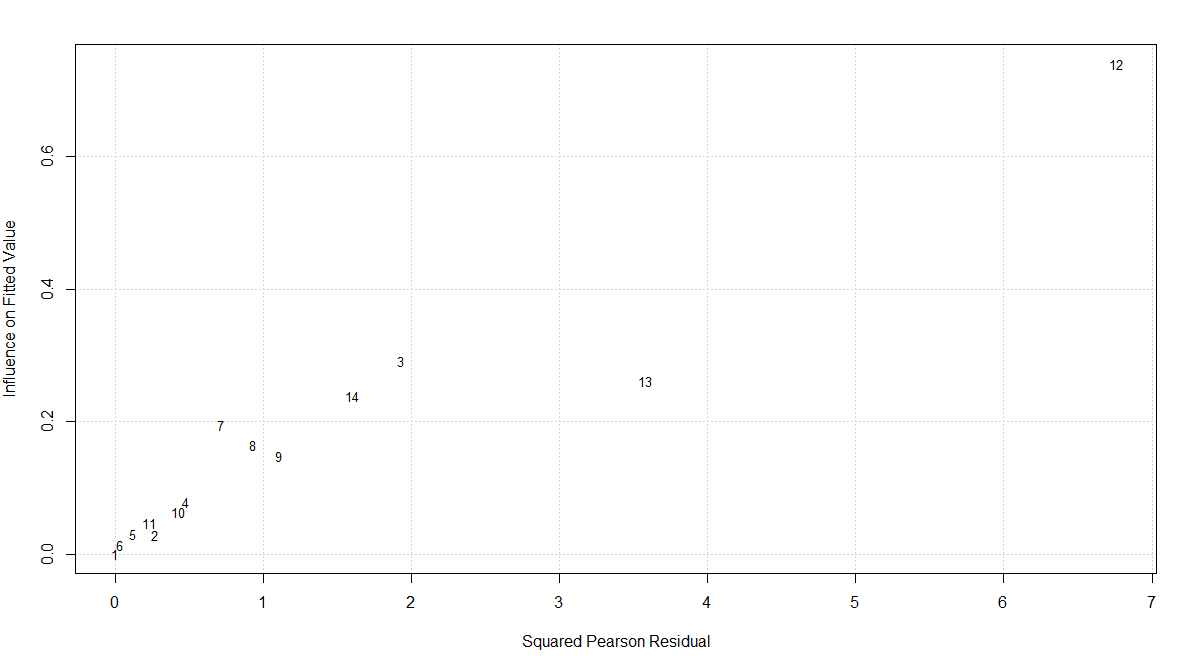


Fig. A5. Baujat plot of the studies for PFS at 6 months (first-line patients).

**Sensitivity meta-analysis**

In this subsection, we perform a sensitivity analysis removing the study of Blay 2014: Trabectedin from the database and re-fitting a random effects model.

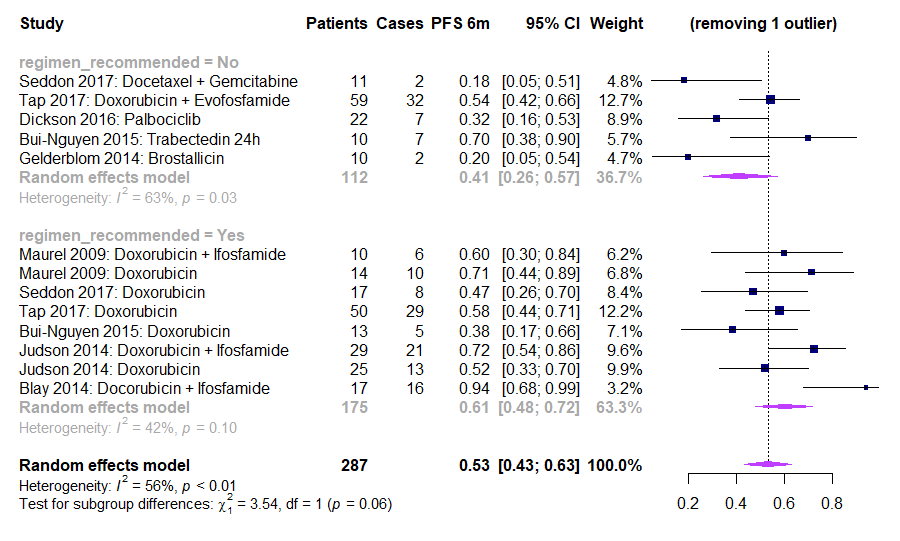


Fig. A6. Forest plot of LPS for PFS at 6 months excluding ‘Blay 2014: Trabectedin' (first-line patients).

Overall effect size of the meta-analysis decreased to 0.53 (0.43-0.63) from 0.56 (0.45-0.67). Overall heterogeneity dropped slightly from 63% (p < 0.01) to 56% (p < 0.01). The study was removed from the non-recommended treatments. The estimate of the subgroup has been recalculated to 0.41 (0.26 – 0.57) from 0.48 (0.31 – 0.66). Heterogeneity was decreased from 76% (p < 0.01) to 63% (p = 0.03). Test of moderator with drugs grouped as recommended or not is not significant (p = 0.06).

**Meta-regressions**

|  |  |
| --- | --- |
| **Moderator** | **Dichotomization** (first group reference, count of studies in parenthesis) |
| Phase of the trial | 2 (4), greater than 2 (10) |
| Study design | Randomized (13), non-randomized (1) |
| Year of activation | 2000-2010 (9), 2011-2015 (5) |
| Sample size | 10-24 (10), >24 (4) |

Again, there is only two non-randomized studies. Thus, we did not consider study design as a moderator.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Moderator**  **(univariate analysis)** | **Estimated PFS (vs ref level)** | **Test for moderator**  **P-value** | **Residual heterogeneity (I2)** | **Test for residual heterogeneity**  **P-value** |
| Phase of the trial:  greater than 2 | 0.60 (vs 0.46) | 0.28 | 63.68% | < 0.01 |
| Year of activation:  2011-2015 | 0.50 (vs 0.60) | 0.39 | 64.73% | < 0.01 |
| Sample size: >24 | 0.59 (vs 0.54) | 0.67 | 64.75% | < 0.01 |

For studies of phase greater than 2 the effect size is increased. The omnibus test for the effect of the moderator provided a p-value of 0.28, not statistically significant. Residual heterogeneity is high (I2 = 63.68%, p < 0.01).

A year of activation between 2011 and 2015 has a negative effect on the proportion of PFS against a study activated between 2000-2010, but not statistically significant (p = 0.39). Residual heterogeneity is high (I2 = 64.73%, p < 0.01).

A sample size above 24 patients has a positive effect (0.21 units added to the intercept) on the observed proportion, however not significant (p = 0.67). Again, residual heterogeneity is high (I2 = 64.75%, p < 0.01).

We did not perform any multivariate moderator analysis as none of these moderators was statistically significant and residual heterogeneity was high.

### **Analysis for pre-treated patients at 3 months**

Our database consists of 17 trials, corresponding to 19 different therapeutic combinations (19 lines in long format).

**Risk of bias assessment**

We examine the risk of publication bias with a contour-enhanced funnel plot and formal diagnostic tests.

In figure 9, studies are symmetrically distributed and inside the borders. There is no evidence of publication bias. The rank correlation test produced a p-value of 0.89 and the Egger’s regression test a p-value of 0.80 using a mixed-effects meta-regression model. There is no indication of publication from the formal examination diagnostics.

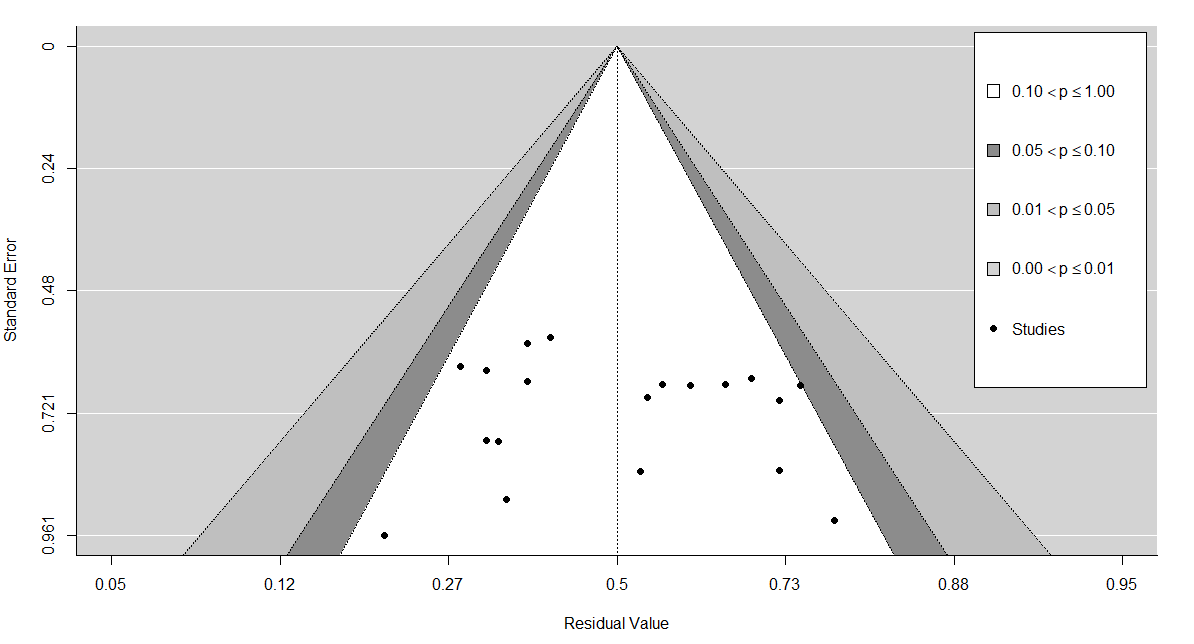


Fig. A7. Contour enhanced funnel plot of LPS for PFS at 3 months (pre-treated patients).

**Diagnostics for influential studies**

Examination of the externally standardized residuals shows that study in line 16 ‘Samuels 2017: Pazopanib’ [9] has the highest z-value of 1.72 so it is potentially influential for the meta-analysis.

The Baujat plot in fig. 11 projects line with number 10 on the upper right corner. ‘Samuels 2017: Pazopanib’ seems to contribute to the highest amount of heterogeneity (based on squared Pearson residuals) and to be the third most influential (fitted values).

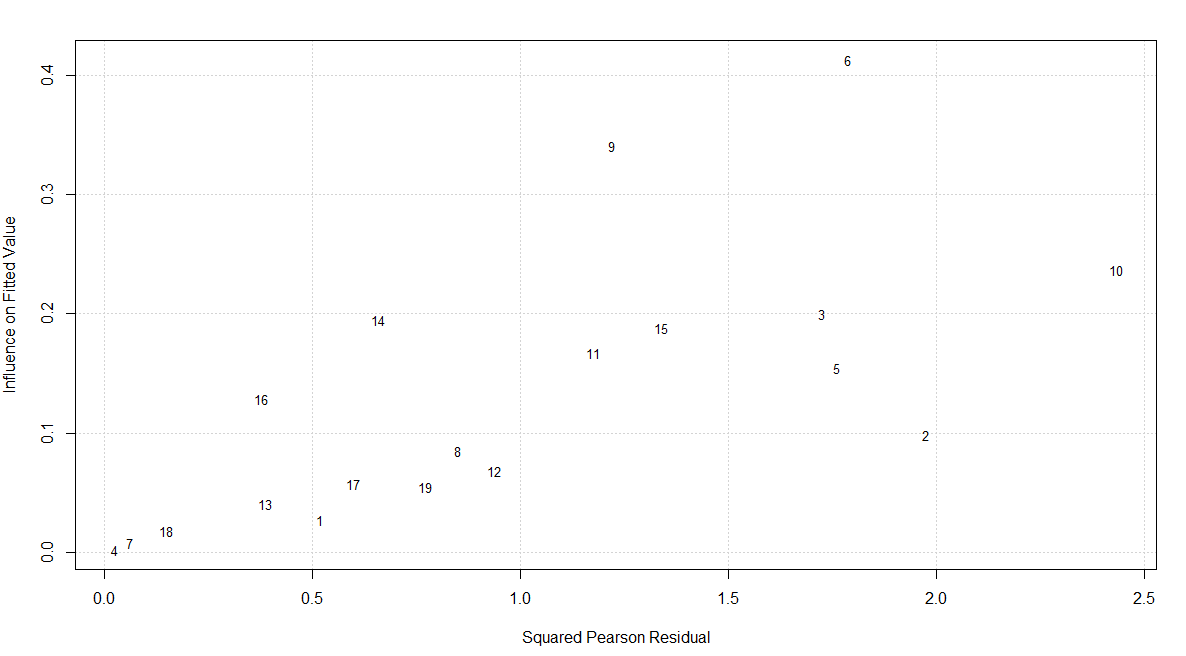


Fig. A8. Baujat plot of the LPS studies for PFS at 3 months (pre-treated patients).

**Sensitivity meta-analysis**

We perform a sensitivity meta-analysis to focus on the impact of ‘Samuels 2017: Pazopanib’ for the random-effects model.

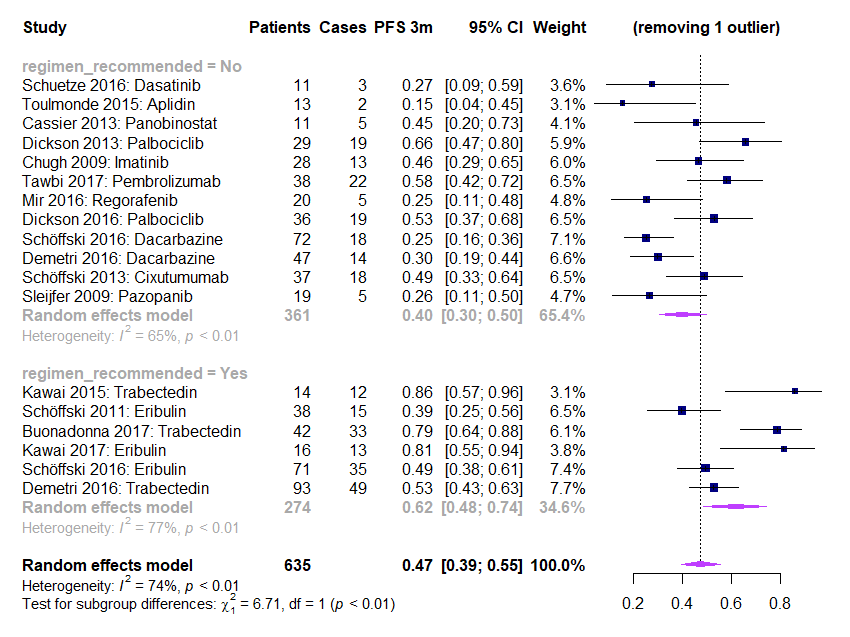


Fig. A9. Forest plot of LPS for PFS at 3 months excluding ‘Samuels 2017: Pazopanib’ (pre-treated patients).

There is a slight decrease in the overall effect size from 0.49 (0.40 – 0.57) to 0.47 (0.39 – 0.55). Overall heterogeneity was barely reduced to 74% p < 0.01 (from 75% p < 0.01). Study was removed from the non-recommended group of drugs. The recalculated estimate for this group is 0.40 (0.30 – 0.50) from 0.42 (0.33 – 0.52). Heterogeneity of the non-recommended subgroup dropped from 71% (p < 0.01) to 65% (p < 0.01). Based on the omnibus chi-square test for moderators, there is still statistically significant difference between the subgroups (p < 0.01). Findings indicate that the excluded study is not that much influential.

**Meta-regressions**

|  |  |
| --- | --- |
| **Moderator** | **Dichotomization** (first group reference, count of studies in parenthesis) |
| Phase of the trial | 2 (14), greater than 2 (5) |
| Study design | Randomized (6), non-randomized (13) |
| Year of activation | 2000-2010 (7), 2011-2015 (12) |
| Sample size | 10-24 (7), >24 (12) |

Potential moderators were examined for the PFS at 3 months (pre-treated patients). All subgroup analyses were conducted with a mixed-effects model. A random-effects model was used to combine the study effects and the fixed effects model was used to test whether these effects vary significantly across the subgroups.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Moderator**  **(univariate analysis)** | **Estimated PFS (vs ref level)** | **Test for moderator**  **P-value** | **Residual heterogeneity (I2)** | **Test for residual heterogeneity**  **P-value** |
| Phase of the trial:  greater than 2 | 0.47 (vs 0.50) | 0.76 | 75.95% | < 0.01 |
| Study design:  non-randomized | 0.52 (vs 0.42) | 0.24 | 73.24% | < 0.01 |
| Year of activation:  2011-2015 | 0.51 (vs 0.44) | 0.40 | 76.18% | < 0.01 |
| Sample size: >24 | 0.51 (vs 0.42) | 0.39 | 75.80% | < 0.01 |

Studies of phase greater than 2 decrease the effect size. The omnibus test for the effect of the moderator based on a chi-square distribution with 1 df provided a p-value of 0.76 (not significant). Residual heterogeneity is high (I2 = 75.95%, p < 0.01). This moderator is not able to explain part of the heterogeneity.

For the pre-treated population, a non-randomized design increases the effect size at 3 months. There is no difference between randomized and non-randomized trials (p = 0.24). Residual heterogeneity is high (73.24%, p < 0.01).

A year of activation between 2011-2015 has a positive effect on the estimate. However, there no statistically significant difference between the 2 subgroups (2011-2015 against 2000-2010) as p = 0.40. Residual heterogeneity is high (75.80%, p < 0.01).

A sample size above 24 patients has a slightly positive effect on the observed proportion versus a sample size of 24 or less. This effect is not statistically significant (p = 0.39). Residual heterogeneity is high (I2 = 75.80%, p < 0.01).

No moderators were found significant. Consequently, no multivariate moderator analysis was performed.

### **Analysis for pre-treated patients at 6 months**

For this case, the EORTC database consists of 17 trials and 19 treatment monotherapies or combinations. Here, PFS at 6 months could not be retrieved for ‘Schöffski 2013: Cixutumumab’ [25] but was available for ‘Chawla 2011: Ridaforolimus’ [26] (which was unavailable for PFS at 3 months).

**Risk of bias assessment**

The contour enhanced funnel plot does not show any study outside the borders. However, there is asymmetry in the distribution of the trials especially for larger standard errors. There is indication of publication bias (could also be citation bias or heterogeneity in the trials).

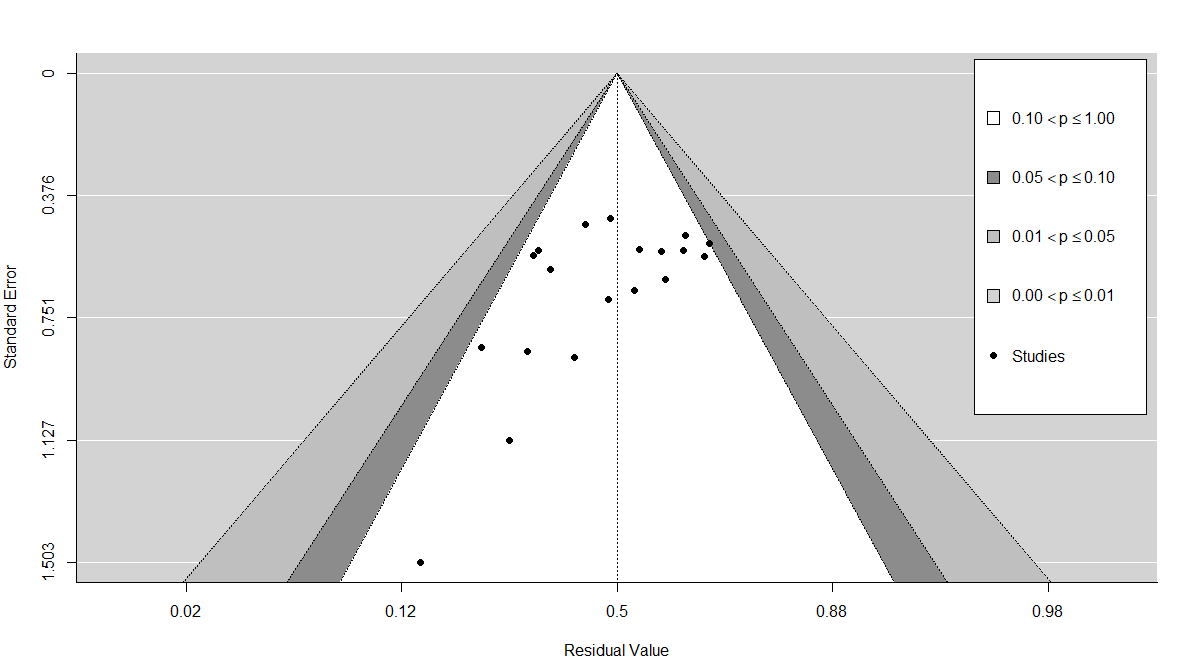


Fig. A10. Contour enhanced funnel plot of LPS for PFS at 6 months (pre-treated patients).

We also performed the 2 formal tests. The rank correlation test produced a p-value of 0.02 and the Egger’s regression test also a p-value of 0.02. There is indication of funnel plot asymmetry from the formal tests which could imply publication bias.

**Diagnostics for influential studies**

The externally standardized residuals of the studies were calculated. Values > 2 or 3 may indicate an influential outlier for the meta-analysis. Exclusion of the most influential study from the database may lead to significant changes in the fitted meta-analytic model. Line 10 with study ‘Samuels 2017: Pazopanib’ [9] has the highest z-value of 1.71.

The baujat plot (fig. 15) projects this study to the upper right corner. This study contributes the most to heterogeneity (squared Pearson residuals) and it is the fourth most influential based on the fitted values on the y-axis.

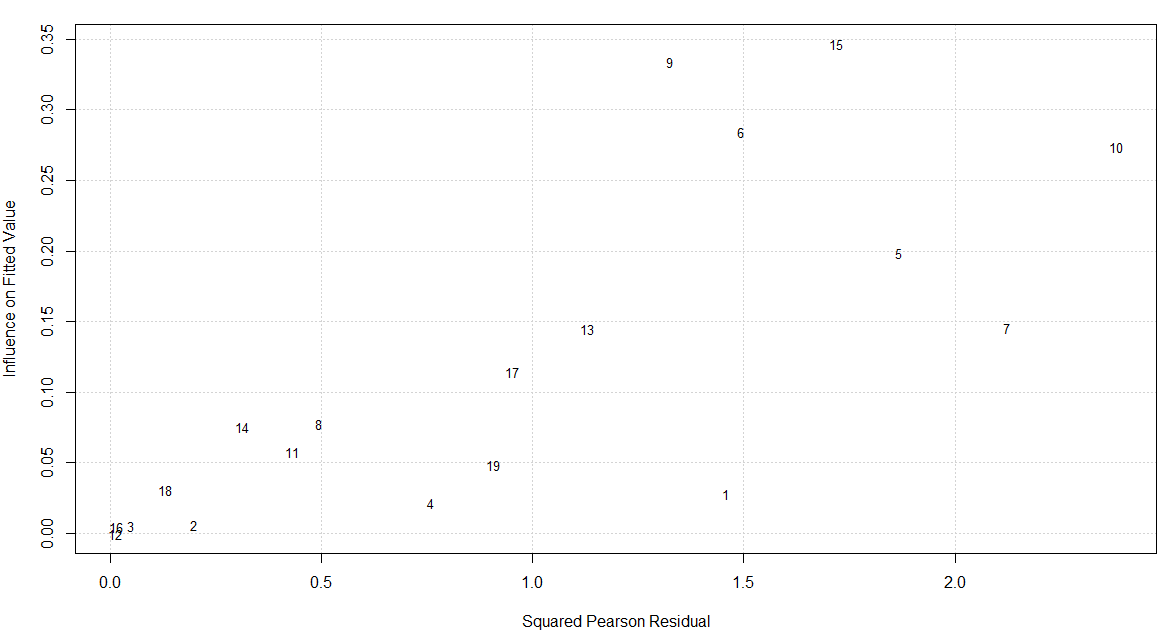


Fig. A11. Baujat plot of the LPS studies for PFS at 6 months (pre-treated patients).

**Sensitivity meta-analysis**

To investigate the effect of the aforementioned study, we performed a sensitivity meta-analysis removing it from the database. We fit a DerSimonian and Laird random-effects model (forest plot in fig. 16). The overall effect size slightly decreases to 0.27 (0.22 – 0.33) from 0.28 (0.22 – 0.34). Overall heterogeneity slightly increased from 66% to 67% (p < 0.01). Study was removed from the non-recommended group of drugs. The estimated effect size of this subgroup decreased to 0.20 (0.14 – 0.26) from 0.21 (0.16 – 0.28). Heterogeneity of the subgroup decreased from 58% (p < 0.01) to 51% (p = 0.02). Based on the omnibus chi-square test for moderators, there is significant difference between the subgroups (p < 0.01). Note that the study of line 1 ‘Schuetze 2016: Dasatinib’ [27] has 0/11 patients progression-free and alive at 6 months but it is not influential because of its small weight (1.0%).

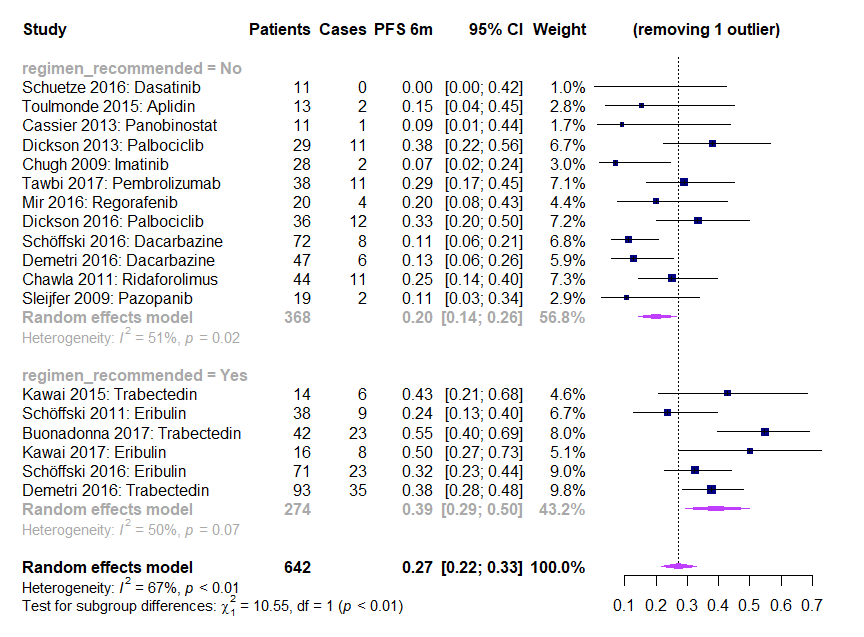


Fig. A12. Forest plot of LPS for PFS at 6 months excluding ‘Samuels 2017: Pazopanib (pre-treated patients).

**Meta-regressions**

|  |  |
| --- | --- |
| **Moderator** | **Dichotomization** (first group reference, count of studies in parenthesis) |
| Phase of the trial | 2 (14), greater than 2 (5) |
| Study design | Randomized (6), non-randomized (13) |
| Year of activation | 2000-2010 (7), 2011-2015 (12) |
| Sample size | 10-24 (7), >24 (12) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Moderator**  **(univariate analysis)** | **Estimated PFS (vs ref level)** | **Test for moderator**  **P-value** | **Residual heterogeneity (I2)** | **Test for residual heterogeneity**  **P-value** |
| Phase of the trial:  greater than 2 | 0.28 (vs 0.26) | 0.80 | 67.95% | < 0.01 |
| Study design:  non-randomized | 0.28 (vs 0.25) | 0.60 | 67.56% | < 0.01 |
| Year of activation:  2011-2015 | 0.31 (vs 0.19) | 0.09 | 64.99% | < 0.01 |
| Sample size: >24 | 0.28 (vs 0.23) | 0.53 | 67.93% | < 0.01 |

Studies of phase greater than 2 (phase 2|3, phase 3, phase 4) increase the effect size. The omnibus test for the effect of the moderator based on a chi-square distribution with 1 degree of freedom provided a p-value of 0.80 (not significant). Residual heterogeneity is high (I2 = 67.95%, p < 0.01).

A non-randomized study versus randomized increases the effect size. The difference between the groups is statistically not significant (p-value = 0.60). Residual heterogeneity is high (I2 = 67.56%, p < 0.01).

A study activated between 2011–2015 versus 2000–2010 increases the effect size. This increase is not significant (p = 0.09). Residual heterogeneity is high (I2 = 67.93%, p < 0.01).

A sample size > 24 versus less than 24 increases the estimate. Based on the test for moderators this difference is not significant (p = 0.53). Residual heterogeneity is high (I2 = 67.93%, p < 0.01).

As there were no significant moderators in the univariate analyses, we did not perform multivariate moderator analysis.

* 1. **Statistical inference for synovial sarcoma (SS)**

We present at separate subsections the risk of bias assessment, corresponding meta-analyses, diagnostics for influential studies, and sensitivity meta-analysis for the studies available. Meta-regressions are not performed here due to the limited number of studies for first-line patients.

### **Analysis for first-line patients at 3 months**

Our database consists of 3 trials on first-line treatment, corresponding to 5 different therapeutic combinations.

**Risk of bias assessment**

We investigate publication bias based on a funnel plot and 2 formal tests. We show the contour enhanced funnel plot in fig 17. Graph is centered at residual value 0.5 (no effect under the null hypothesis). There are shaped regions for various levels of statistical significance for the points (see the label). There is 1 study in the dark grey zone (0.05 ≤ p ≤ 0.10) on the right side.

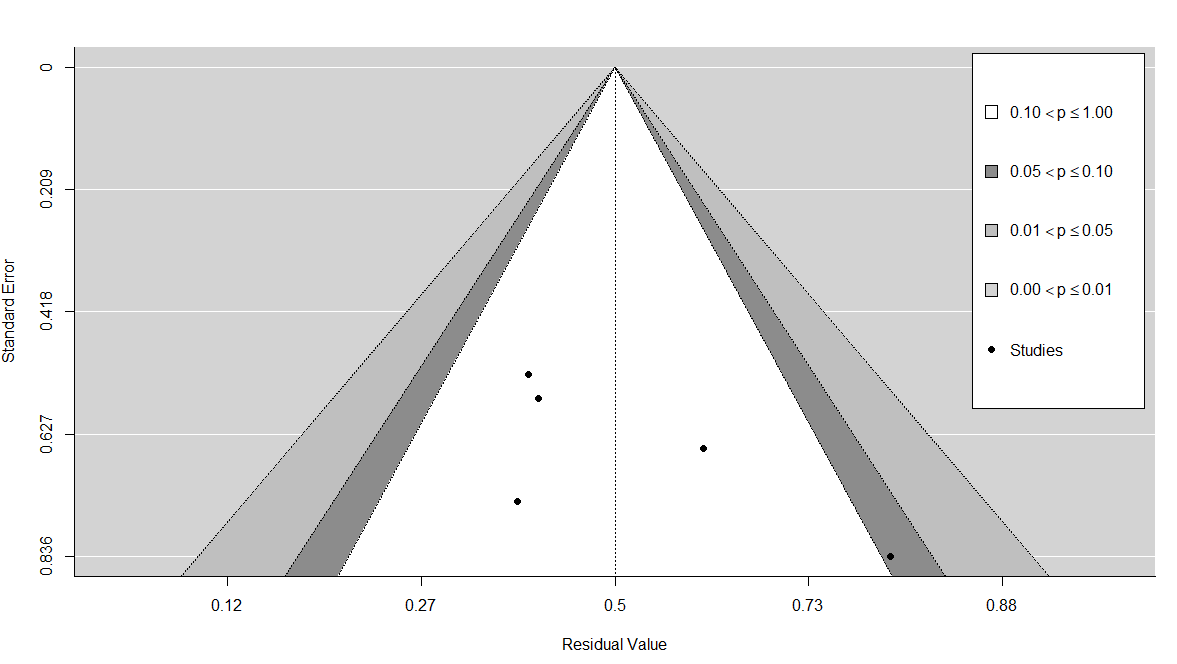


Fig. A13. Contour enhanced funnel plot of SS for PFS at 3 months (first-line patients).

We performed 2 formal tests: (1) the rank correlation test for funnel plot asymmetry, and (2) Egger’s regression test for funnel plot asymmetry. Rank correlation test provided a p-value of 0.08 and Egger’s regression test a p-value of 0.16. Publication bias cannot be excluded as because of the very limited sample size the tests were severely underpowered.

**Diagnostics for influential studies**

We examine the externally standardized residuals of the studies. Z values > 2 or 3 show an influential outlier (its exclusion may lead to significant changes in the fitted meta-analytic model). Study line 3 of Judson 2014: Doxorubicin+Ifosfamide [28] had a value of 2.37.

We used the Baujat plot (figure 19) as a diagnostic tool to detect sources of heterogeneity and potentially influential studies. Study with number 3 (Judson 2014: Doxorubicin+Ifosfamide) is projected on the right upper corner (contributes to heterogeneity and is the most influential).

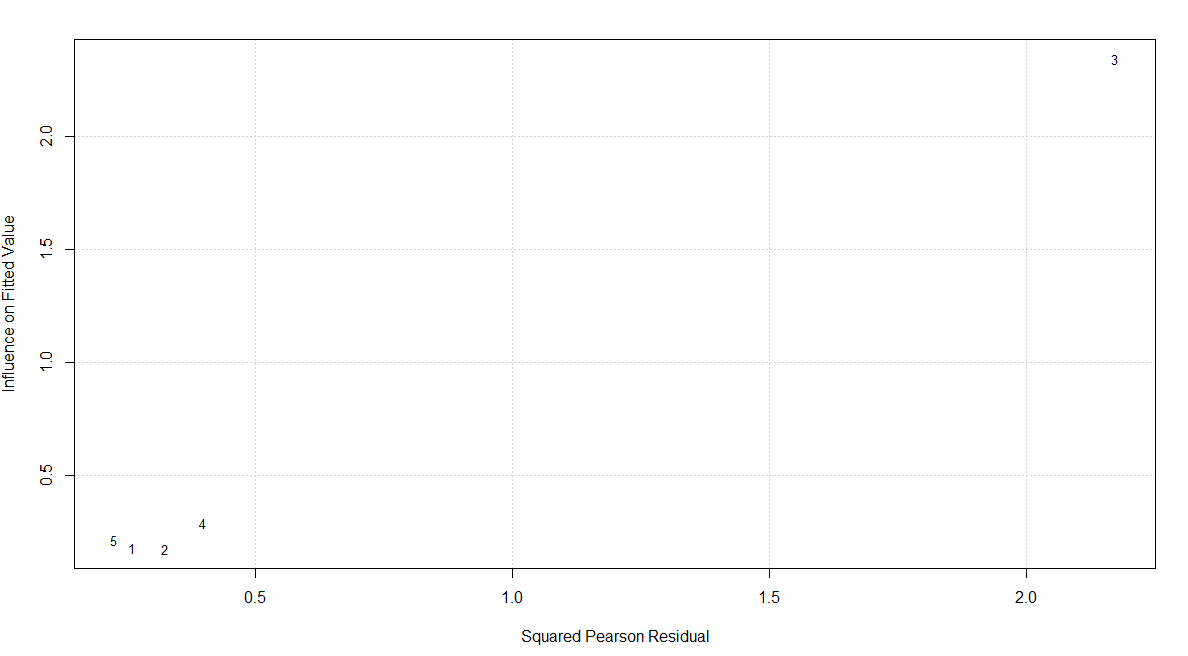


Fig. A14. Baujat plot of the SS studies for PFS at 3 months (first-line patients).

**Sensitivity meta-analysis**

To investigate the effect of ‘Judson 2014: Doxorubicin+Ifosfamide’ which was found as influential in previous section, we performed a sensitivity analysis removing this treatment arm from the analysis and re-fitting the random-effects model.

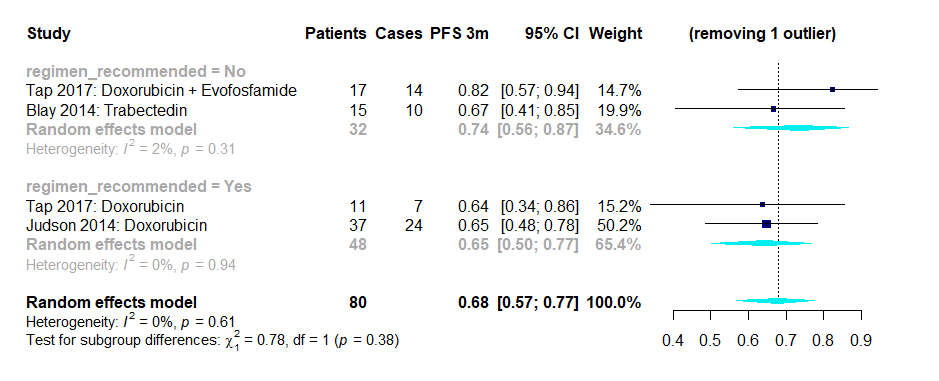


Fig. A15. Forest plot of SS for PFS at 3 months excluding ‘Judson 2014: Doxorubicin+Ifosfamide' (first-line patients).

There is a decrease in the overall effect size of the meta-analysis from 0.74 (0.58 – 0.86) to 0.68 (0.57 – 0.77). Overall heterogeneity dropped significantly from 41% (p=0.15) to 0% (p = 0.61). This treatment arm was removed from the recommended treatment regimens. The estimate of the subgroup has been recalculated to 0.65 (0.50 – 0.77) from 0.74 (0.53 – 0.88). There is a big decrease in the group heterogeneity to 0%, p = 0.94 (from 64%, p = 0.06). Again, the difference between a regimen recommended versus a non-recommended is not statistically significant based on moderator’s test (p = 0.38). Findings indicate that ‘Judson 2014: Doxorubicin+Ifosfamide' is an influential study. This sensitivity meta-analysis indicates that the results of the 3-month PFS for the SS first-line patients are not robust which makes sense given the very limited number of available studies (5 rows from 3 trials).

### **Analysis for first-line patients at 6 months**

Our database consists of 3 trials covering 5 different therapeutic combinations.

**Risk of bias assessment**

Again, we investigate publication bias with a funnel plot and 2 formal tests. The contour enhanced funnel plot in fig. 21 did not detect studies outside the triangular borders. Publication bias cannot be excluded as it is challenging to visually inspect asymmetry because of the limited number of studies.

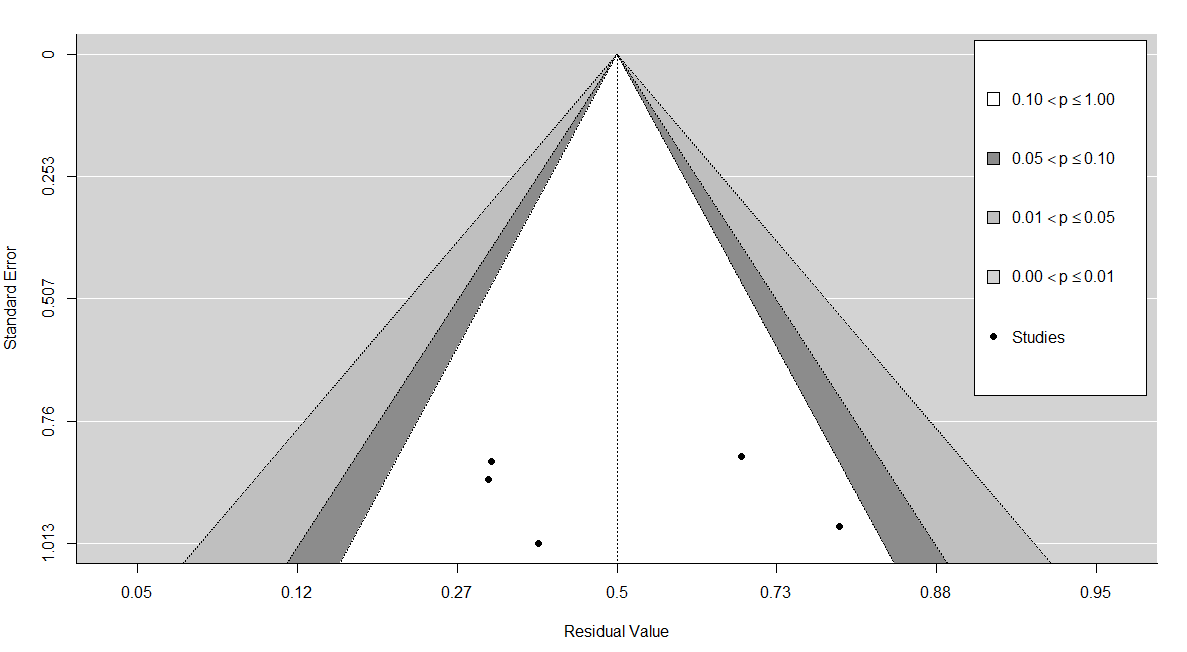


Fig. A16. Contour enhanced funnel plot of SS for PFS at 6 months (first-line patients).

Hence, we examine the risk of bias more formally. The p-values from the rank correlation and the Egger’s regression tests were 0.82 and 0.60, respectively. There is no indication of publication bias here. Note that these results might have been affected by the very limited number of studies.

**Diagnostics for influential studies**

Examination of the externally standardized residuals shows that study in line 3 ‘Judson 2014: Doxorubicin + Ifosfamide’ [28] has the highest z-value of 2.22, so it is potentially influential for the meta-analysis.

The Baujat plot in fig. 23 projects line with number 3 on the upper right corner. ‘Judson 2014: Doxorubicin + Ifosfamide’ seems to contribute to the highest amount of heterogeneity (based on squared Pearson residuals) and to be the most influential (fitted values).

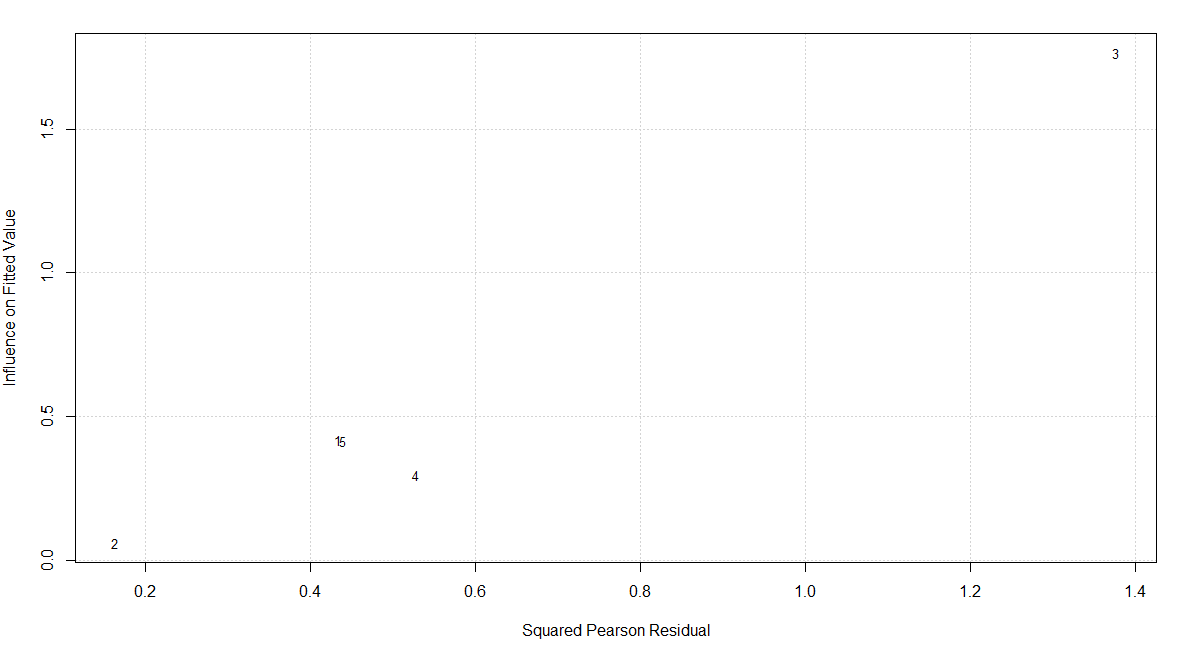


Fig. A17. Baujat plot of the SS studies for PFS at 6 months (first-line patients).

**Sensitivity meta-analysis**

In this subsection, we perform a sensitivity analysis removing the study of ‘Judson 2014: Doxorubicin + Ifosfamide’ from the database and re-fitting a random effects model.

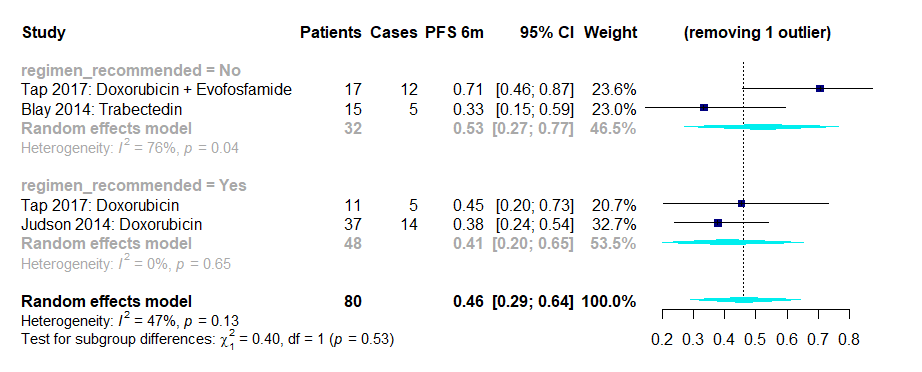


Fig. A18. Forest plot of SS for PFS at 6 months excluding ‘Judson 2014: Doxorubicin + Ifosfamide’ (first-line patients).

Overall effect size of the meta-analysis decreased to 0.46 (0.29-0.64) from 0.56 (0.31 – 0.78). Overall heterogeneity dropped from 75% (p < 0.01) to 47% (p = 0.13). The study was removed from the recommended treatments. The estimate of the subgroup has been recalculated to 0.41 (0.20 – 0.65) from 0.58 (0.27 – 0.84). Heterogeneity was decreased significantly from 83% (p<0.01) to 0% (p = 0.65). Test of moderator with drugs grouped as recommended or not was not significant (p = 0.53). Again, results were not robust which can be attributed to the small number of studies (5 rows in long format).

### **Analysis for pre-treated patients at 3 months**

Our database consists of 10 trials covering 10 different therapeutic combinations. Two of these trials were randomized (van der Graaf 2012 [29], Mir 2016 [30]). The control arms contained placebo treatment and were excluded from the analysis.

**Risk of bias assessment**

We examine the risk of publication bias with a contour-enhanced funnel plot and formal diagnostic tests. In figure 25, there is one studies in the dark grey zone (right side, 0.05 < p ≤ 0.10) and one more in the grey zone (0.01 < p ≤ 0.05). There seems to be some asymmetry in the study distribution. The rank correlation test produced a p-value of 0.60 and the Egger’s regression test a p-value of 0.19. There is no indication of high risk for publication bias from the formal examination diagnostics.

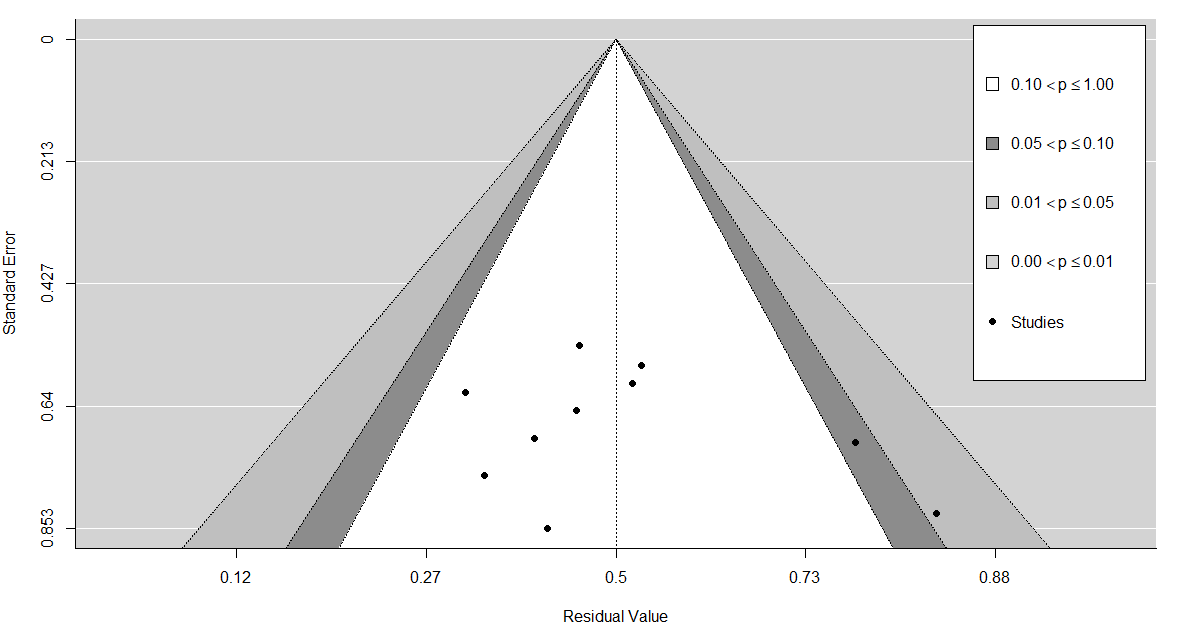


Fig. A19. Contour enhanced funnel plot of SS for PFS at 3 months (pre-treated patients).

**Diagnostics for influential studies**

Examination of the externally standardized residuals shows that studies in line 5 ‘Mir 2016: Regorafenib’ [30] and line 6 ‘Robbins 2015: Cyclosphosphamide + Fludarabine + TCR transduced cells’ [31] have the highest z-value of 2.34 and 2.23, respectively. Both studies are potentially influential for the meta-analysis. The residual plots (not shown) indicate line 6 more influential.

The Baujat plot in fig. 27 projects line with number 6 on the upper right corner. ‘Robbins 2015: Cyclosphosphamide + Fludarabine + TCR transduced cells’ seems to contribute to the second highest amount of heterogeneity (based on squared Pearson residuals) and to be the most influential (fitted values).

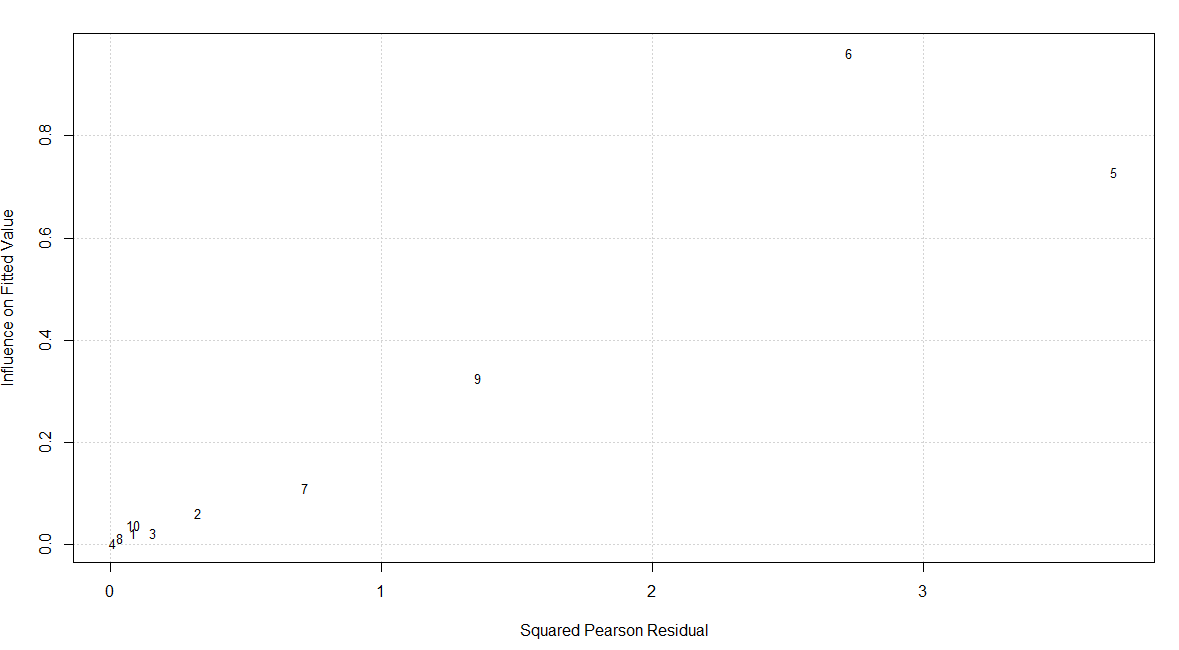


Fig. A20. Baujat plot of the SS studies for PFS at 3 months (pre-treated patients).

**Sensitivity meta-analysis**

We perform a sensitivity meta-analysis to focus on the impact of ‘Robbins 2015: Cyclosphosphamide + Fludarabine + TCR transduced cells’ for the random-effects model.

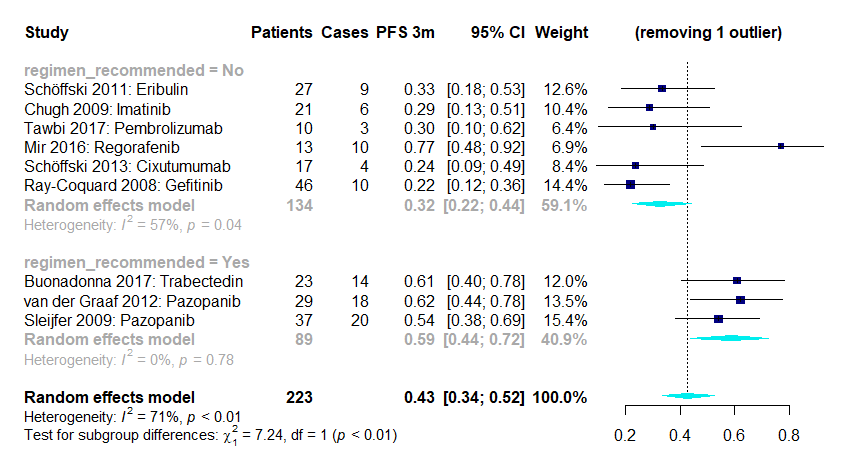


Fig. A21. Forest plot of SS for PFS at 3 months excluding ‘Robbins 2015: Cyclosphosphamide + Fludarabine + TCR transduced cells’ (pre-treated patients).

There is a decrease in the overall effect size from 0.45 (0.34 – 0.57) to 0.43 (0.34 – 0.52). The overall heterogeneity is just reduced to 71% p < 0.01 (from 72% p < 0.01). Study was removed from the non-recommended group of drugs. The recalculated estimate for this group is 0.32 (0.22 – 0.44) from 0.38 (0.26 – 0.52). Based on the omnibus chi-square test for moderators, this time there is statistically significant difference between the subgroups (p < 0.01). Findings indicate that the excluded study might be influential for the PFS estimates. Results are not very robust.

**Meta-regressions**

|  |  |
| --- | --- |
| **Moderator** | **Dichotomization** (first group reference, count of studies in parenthesis) |
| Phase of the trial | 2 (8), greater than 2 (2) |
| Study design | Randomized (2), non-randomized (8) |
| Year of activation | 2000-2008 (7), 2009-2015 (3) |
| Sample size | 10-21 (5), >21 (5) |

Potential moderators were examined for the PFS at 3 months (pre-treated patients). All subgroup analyses were conducted with a mixed-effects model. A random-effects model was used to combine the study effects and the fixed effects model was used to test whether these effects vary significantly across the subgroups. We only performed analyses for moderators including more than two study rows; here year of activation and sample size.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Moderator**  **(univariate analysis)** | **Estimated PFS (vs ref level)** | **Test for moderator**  **P-value** | **Residual heterogeneity (I2)** | **Test for residual heterogeneity**  **P-value** |
| Year of activation:  2009-2015 | 0.57 (vs 0.41) | 0.28 | 71.56% | < 0.01 |
| Sample size: >21 | 0.46 (vs 0.45) | 0.97 | 74.83 | < 0.01 |

A year of activation between 2009-2015 has a positive effect on the estimate. However, there no statistically significant difference between the 2 subgroups (2009-2015 against 2000-2008) as p = 0.28. Residual heterogeneity is high (71.56%, p < 0.01).

A sample size above 21 patients has a slightly positive effect on the observed proportion versus a sample size of 21 or less. This effect is not statistically significant (p = 0.97). Residual heterogeneity is high (I2 = 74.83%, p < 0.01).

No moderators were found significant. Consequently, no multivariate moderator analysis was performed.

### **Analysis for pre-treated patients at 6 months**

The EORTC database included 9 trials covering 9 different therapeutic combinations. Information for PFS at 6 months could not be retrieved for ‘Schöffski 2013: Cixutumumab’ [25].

**Risk of bias assessment**

The contour enhanced funnel plot shows no study with p ≤ 0.10. However, there seems to be some asymmetry between the studies on the right and on the left side of the plot for higher standard error values.

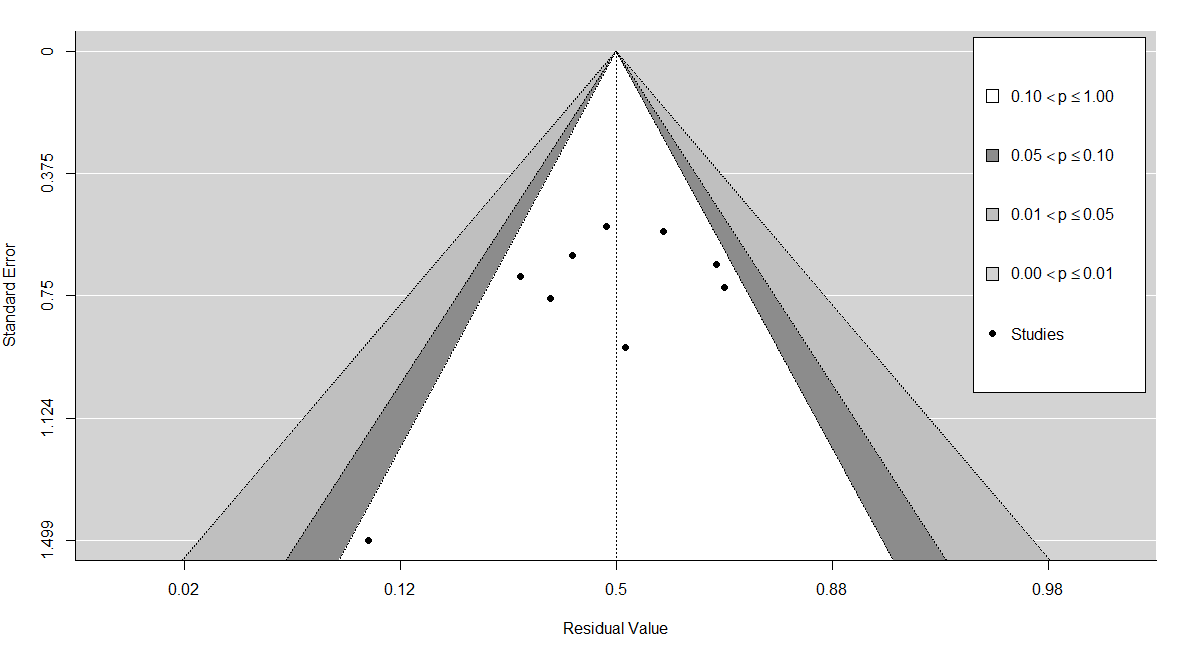


Fig. A22. Contour enhanced funnel plot of SS for PFS at 6 months (pre-treated patients).

We also performed the 2 formal tests. The rank correlation test produced a p-value of 0.12 and the Egger’s regression test a p-value of 0.09. There is no indication of publication bias based on the formal tests.

**Diagnostics for influential studies**

The externally standardized residuals of the studies were calculated. Line 2 and 6 of studies ‘Chugh 2009: Imatinib’ [32] and ‘Robbins 2015: Cyclosphosphamide + Fludarabine + TCR transduced cells’ [31] have the highest z-values of -1.56 and 1.55, respectively.

The baujat plot (fig. 31) projects this study to the upper right side (number 6). This study is the 2nd most influential based on the fitted values on the y-axis and contributes the 2nd most to heterogeneity looking at the squared Pearson residuals (x-axis).

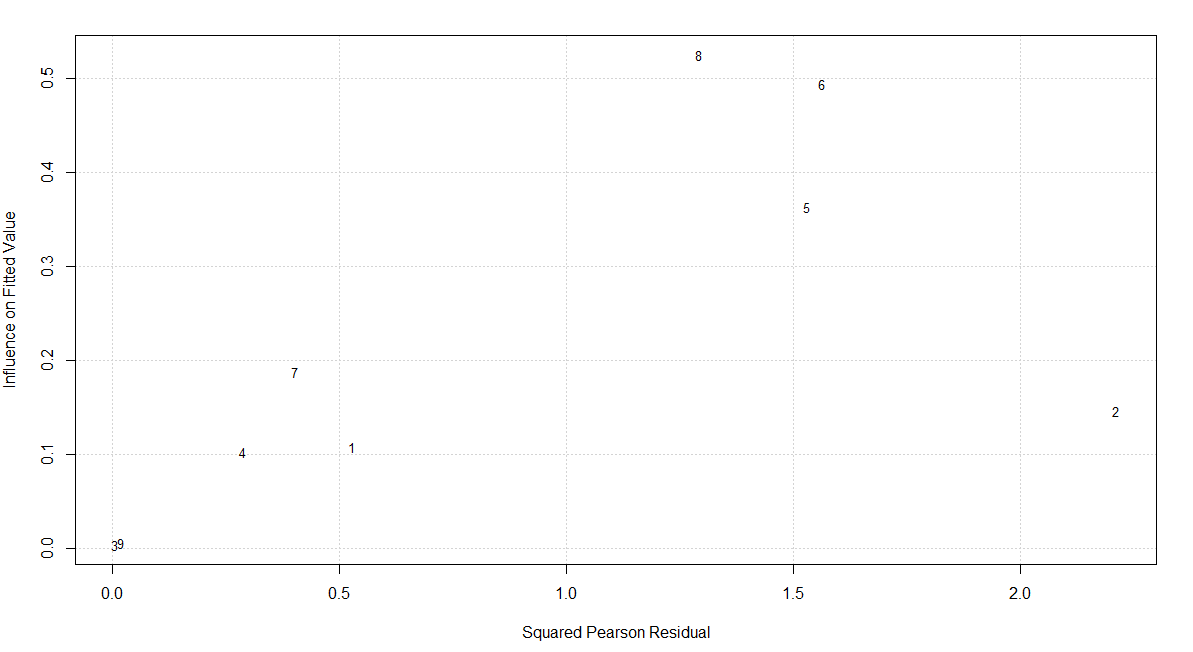


Fig. A23. Baujat plot of the SS studies for PFS at 6 months (pre-treated patients).

**Sensitivity meta-analysis**

We performed a sensitivity meta-analysis removing ‘Robbins 2015: Cyclosphosphamide + Fludarabine + TCR transduced cells’ from the database. We fit a DerSimonian and Laird random-effects model (forest plot in fig. 32). The overall effect size slightly decreased to 0.24 (0.16 – 0.34) from 0.25 (0.16 – 0.36). Overall heterogeneity is 64% (p < 0.01). Study was removed from the non-recommended group of drugs. The estimated effect size of this subgroup decreased to 0.15 (0.08 – 0.27) from 0.19 (0.10 – 0.32). Based on the omnibus chi-square test for moderators, the difference between the subgroups is now statistically significant (p = 0.04). Findings are more robust compared to 3 months. Nevertheless, the subgroups of drugs turn to be statistically different after the exclusion of the Robbins study (similar to PFS at 3 months).

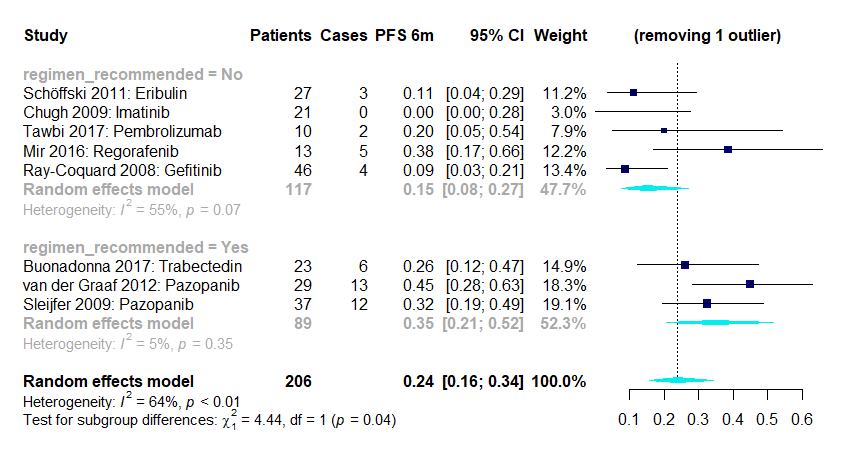


Fig. A24. Forest plot of SS for PFS at 6 months excluding ‘Robbins 2015: Cyclosphosphamide + Fludarabine + TCR transduced cells’ (pre-treated patients).

**Meta-regressions**

|  |  |
| --- | --- |
| **Moderator** | **Dichotomization** (first group reference, count of studies in parenthesis) |
| Phase of the trial | 2 (7), greater than 2 (2) |
| Study design | Randomized (2), non-randomized (7) |
| Year of activation | 2000-2008 (6), 2009-2015 (3) |
| Sample size | 10-21 (4), >21 (5) |

Similarly, we perform univariate meta-regressions only for the year of activation and sample size.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Moderator**  **(univariate analysis)** | **Estimated PFS (vs ref level)** | **Test for moderator**  **P-value** | **Residual heterogeneity (I2)** | **Test for residual heterogeneity**  **P-value** |
| Year of activation:  2009-2015 | 0.28 (vs 0.22) | 0.63 | 65.91% | < 0.01 |
| Sample size: >21 | 0.23 (vs 0.26) | 0.80 | 65.56% | < 0.01 |

A study activated between 2009–2015 versus 2000–2008 increases the effect size. This increase is not significant (p = 0.63). Residual heterogeneity is high (I2 = 65.91%, p < 0.01).

A sample size > 21 versus less than 21 decreases the estimate. Based on the test for moderators this difference is not significant (p = 0.80). Residual heterogeneity is high (I2 = 65.56%, p < 0.01).

As there were no significant moderators in the univariate analyses, we did not perform any multivariate moderator analysis.

# References

[1] Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. J Clin Epidemiol 2009;62:e1--e34. https://doi.org/10.1016/j.jclinepi.2009.06.006.

[2] Maurel J, López-Pousa A, De Las Peñas R, Fra J, Martín J, Cruz J, et al. Efficacy of sequential high-dose doxorubicin and ifosfamide compared with standard-dose doxorubicin in patients with advanced soft tissue sarcoma: An open-label randomized phase II study of the Spanish group for research on sarcomas. J Clin Oncol 2009;27:1893–8. https://doi.org/10.1200/JCO.2008.19.2930.

[3] Bui-Nguyen B, Butrynski JE, Penel N, Blay JY, Isambert N, Milhem M, et al. A phase IIb multicentre study comparing the efficacy of trabectedin to doxorubicin in patients with advanced or metastatic untreated soft tissue sarcoma: The TRUSTS trial. Eur J Cancer 2015;51:1312–20. https://doi.org/10.1016/j.ejca.2015.03.023.

[4] Gelderblom H, Blay JY, Seddon BM, Leahy M, Ray-Coquard I, Sleijfer S, et al. Brostallicin versus doxorubicin as first-line chemotherapy in patients with advanced or metastatic soft tissue sarcoma: An European Organisation for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group randomised phase II and pharmacogeneti. Eur J Cancer 2014;50:388–96. https://doi.org/10.1016/j.ejca.2013.10.002.

[5] Kawai A, Araki N, Sugiura H, Ueda T, Yonemoto T, Takahashi M, et al. Trabectedin monotherapy after standard chemotherapy versus best supportive care in patients with advanced, translocation-related sarcoma: A randomised, open-label, phase 2 study. Lancet Oncol 2015;16:406–16. https://doi.org/10.1016/S1470-2045(15)70098-7.

[6] Dickson MA, Tap WD, Keohan ML, D’Angelo SP, Gounder MM, Antonescu CR, et al. Phase II trial of the CDK4 inhibitor PD0332991 in patients with advanced CDK4-amplified well-differentiated or dedifferentiated liposarcoma. J Clin Oncol 2013;31:2024–8. https://doi.org/10.1200/JCO.2012.46.5476.

[7] Dickson MA, Schwartz GK, Louise Keohan M, D’Angelo SP, Gounder MM, Chi P, et al. Phase 2 Trial of the CDK4 inhibitor Palbociclib (PD0332991) at 125 mg dose in Well-Differentiated or Dedifferentiated Liposarcoma. JAMA Oncol 2016;2:937–40. https://doi.org/10.1001/jamaoncol.2016.0264.

[8] Toulmonde M, Le Cesne A, Piperno-Neumann S, Penel N, Chevreau C, Duffaud F, et al. Aplidin in patients with advanced dedifferentiated liposarcomas: A French Sarcoma Group Single-Arm Phase II study. Ann Oncol 2015;26:1465–70. https://doi.org/10.1093/annonc/mdv195.

[9] Samuels BL, Chawla SP, Somaiah N, Staddon AP, Skubitz KM, Milhem MM, et al. Results of a prospective phase 2 study of pazopanib in patients with advanced intermediate-grade or high-grade liposarcoma. Cancer 2017;123:4640–7. https://doi.org/10.1002/cncr.30926.

[10] Gronchi A, Miah AB, Dei Tos AP, Abecassis N, Bajpai J, Bauer S, et al. Soft tissue and visceral sarcomas: ESMO–EURACAN–GENTURIS Clinical Practice Guidelines for diagnosis, treatment and follow-up. Ann Oncol 2021;32:1348–65. https://doi.org/10.1016/j.annonc.2021.07.006.

[11] Blay JY, Leahy MG, Nguyen BB, Patel SR, Hohenberger P, Santoro A, et al. Randomised phase III trial of trabectedin versus doxorubicin-based chemotherapy as first-line therapy in translocation-related sarcomas. Eur J Cancer 2014;50:1137–47. https://doi.org/10.1016/j.ejca.2014.01.012.

[12] Ray-Coquard I, Le Cesne A, Whelan JS, Schoffski P, Bui BN, Verweij J, et al. A Phase II Study of Gefitinib for Patients with Advanced HER-1 Expressing Synovial Sarcoma Refractory to Doxorubicin-Containing Regimens. Oncologist 2008;13:467–73. https://doi.org/10.1634/theoncologist.2008-0065.

[13] Dersimonian R, Laird N. Meta-Analysis in Clinical Trials. Control Clin Trials 1986;7:177–88.

[14] Dersimonian R, Laird N. Meta-Analysis in Clinical Trials Revisited. Contemp Clin Trials 2015;45:139–45. https://doi.org/10.1016/j.cct.2015.09.002.Meta-Analysis.

[15] Borenstein M, Hedges L V., Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. John Wiley & Sons; 2011.

[16] Viechtbauer W, Cheung MW-L. Outlier and influence diagnostics for meta-analysis. Res Synth Methods 2010;1:112–25. https://doi.org/10.1002/jrsm.11.

[17] Baujat B, Mahé C, Pignon JP, Hill C. A graphical method for exploring heterogeneity in meta-analyses: Application to a meta-analysis of 65 trials. Stat Med 2002;21:2641–52. https://doi.org/10.1002/sim.1221.

[18] Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. J Clin Epidemiol 2001;54:1046–55. https://doi.org/10.1016/S0895-4356(01)00377-8.

[19] Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. J Clin Epidemiol 2008;61:991–6. https://doi.org/10.1016/j.jclinepi.2007.11.010.

[20] Begg CB, Mazumdar M. Operating Characteristics of a Rank Correlation Test for Publication Bias. Biometrics 1994;50:1088–101. https://doi.org/10.2307/2533446.

[21] Egger M, Smith GD, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test measures of funnel plot asymmetry. BMJ 1997;315:629–34. https://doi.org/10.1136/bmj.315.7109.629.

[22] Wang N. How to Conduct a Meta-Analysis of Proportions in R : A Comprehensive Tutorial Conducting Meta-Analyses of Proportions in R. John Jay Coll Crim Justice 2018;1:1–62. https://doi.org/10.13140/RG.2.2.27199.00161.

[23] Sahai H, Ageel MI. The analysis of variance : fixed, random, and mixed models. Springer Science and Business Media; 2012. https://doi.org/https://doi.org/10.1007/978-1-4612-1344-4.

[24] Miller JJ. The inverse of the freeman-tukey double arcsine transformation. Am Stat 1978;32:138–138. https://doi.org/10.1080/00031305.1978.10479283.

[25] Schöffski P, Adkins D, Blay JY, Gil T, Elias AD, Rutkowski P, et al. An open-label, phase 2 study evaluating the efficacy and safety of the anti-IGF-1R antibody cixutumumab in patients with previously treated advanced or metastatic soft-tissue sarcoma or Ewing family of tumours. Eur J Cancer 2013;49:3219–28. https://doi.org/10.1016/j.ejca.2013.06.010.

[26] Chawla SP, Staddon AP, Baker LH, Schuetze SM, Tolcher AW, D’Amato GZ, et al. Phase II study of the mammalian target of rapamycin inhibitor ridaforolimus in patients with advanced bone and soft tissue sarcomas. J Clin Oncol 2012;30:78–84. https://doi.org/10.1200/JCO.2011.35.6329.

[27] Schuetze SM, Wathen JK, Lucas DR, Choy E, Samuels BL, Staddon AP, et al. SARC009: Phase 2 study of dasatinib in patients with previously treated, high-grade, advanced sarcoma. Cancer 2016;122:868–74. https://doi.org/10.1002/cncr.29858.

[28] Judson I, Verweij J, Gelderblom H, Hartmann JT, Schöffski P, Blay JY, et al. Doxorubicin alone versus intensified doxorubicin plus ifosfamide for first-line treatment of advanced or metastatic soft-tissue sarcoma: A randomised controlled phase 3 trial. Lancet Oncol 2014;15:415–23. https://doi.org/10.1016/S1470-2045(14)70063-4.

[29] Van Der Graaf WTA, Blay JY, Chawla SP, Kim DW, Bui-Nguyen B, Casali PG, et al. Pazopanib for metastatic soft-tissue sarcoma (PALETTE): A randomised, double-blind, placebo-controlled phase 3 trial. Lancet 2012;379:1879–86. https://doi.org/10.1016/S0140-6736(12)60651-5.

[30] Mir O, Brodowicz T, Italiano A, Wallet J, Blay JY, Bertucci F, et al. Safety and efficacy of regorafenib in patients with advanced soft tissue sarcoma (REGOSARC): a randomised, double-blind, placebo-controlled, phase 2 trial. Lancet Oncol 2016;17:1732–42. https://doi.org/10.1016/S1470-2045(16)30507-1.

[31] Robbins PF, Kassim SH, Tran TLN, Crystal JS, Morgan RA, Feldman SA, et al. A pilot trial using lymphocytes genetically engineered with an NY-ESO-1-reactive T-cell receptor: Long-term follow-up and correlates with response. Clin Cancer Res 2015;21:1019–27. https://doi.org/10.1158/1078-0432.CCR-14-2708.

[32] Chugh R, Wathen JK, Maki RG, Benjamin RS, Patel SR, Myers PA, et al. Phase II multicenter trial of imatinib in 10 histologic subtypes of sarcoma using a bayesian hierarchical statistical model. J Clin Oncol 2009;27:3148–53. https://doi.org/10.1200/JCO.2008.20.5054.