**Efficacy thresholds for clinical trials with advanced or metastatic leiomyosarcoma patients: A European Organisation for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group meta-analysis based on a literature review for soft-tissue sarcomas**

Georgios Kantidakis, Saskia Litière, Anouk Neven, Marie Vinches, Ian Judson, Patrick Schöffski, 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](#_Toc66885715)

[2 Details about the Leiomyosarcoma meta-analyses 5](#_Toc66885716)

[2.1 Study characteristics 5](#_Toc66885717)

[2.1.1 Studies for all LMS meta-analyses 5](#_Toc66885718)

[2.1.2 Studies for uterine LMS meta-analyses 8](#_Toc66885719)

[2.2 Statistical considerations 11](#_Toc66885720)

[2.2.1 Unit of the analyses 11](#_Toc66885721)

[2.2.2 Meta-analysis methodology 11](#_Toc66885722)

[2.2.3 Hypothesis generating analyses 11](#_Toc66885723)

[2.2.4 Sensitivity meta-analysis 12](#_Toc66885724)

[2.2.5 Risk of bias assessment 12](#_Toc66885725)

[2.2.6 Proportions and transformation function 12](#_Toc66885726)

[2.3 Statistical inference for all LMS 13](#_Toc66885727)

[2.3.1 Analysis for first-line patients at 3 months 13](#_Toc66885728)

[2.3.2 Analysis for first-line patients at 6 months 16](#_Toc66885729)

[2.3.3 Analysis for pre-treated patients at 3 months 19](#_Toc66885730)

[2.3.4 Analysis for pre-treated patients at 6 months 22](#_Toc66885731)

[2.4 Statistical inference for uterine LMS 25](#_Toc66885732)

[2.4.1 Analysis for first-line patients at 3 months 25](#_Toc66885733)

[2.4.2 Analysis for first-line patients at 6 months 28](#_Toc66885734)

[2.4.3 Analysis for pre-treated patients at 3 months 30](#_Toc66885735)

[2.4.4 Analysis for pre-treated patients at 6 months 33](#_Toc66885736)

[3 References 35](#_Toc66885737)

# 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 Leiomyosarcoma meta-analyses

The focus for this meta-analysis is on the most frequent STS subtype appearing in the 159 trials; Leiomyosarcoma (LMS). Overall, more than 100 trials contained LMS as one of the subtypes (LMS, uLMS or soft-tissue LMS etc.).

* 1. **Study characteristics**

### **Studies for all LMS meta-analyses**

In this subsection we provide descriptive statistics regarding the **23 trials** with all LMS included in the meta-analysis, and the corresponding **31 drug regimens** of the long version of the database. For instance, the phase III trial of Tap et al.2 included 2 treatment regimens either doxorubicin monotherapy or doxorubicin in combination with Evofosfamide.

In the table below, year of study activation (year that first patient was recruited) is presented for the 23 trials in the STBSG database. Most common years of activation at trial level were 2010 and 2011 (17.39% each).

|  |  |  |
| --- | --- | --- |
| **Year of study activation** | **N per trial (%)** | **N per drug regimen (%)** |
| 2002 | 2 (8.70%) | 2 (6.45%) |
| 2003 | 1 (4.35%) | 2 (6.45%) |
| 2004 | 1 (4.35%) | 1 (3.23%) |
| 2005 | 1 (4.35%) | 1 (3.23%) |
| 2006 | 3 (13.04%) | 5 (16.03%) |
| 2007 | 2 (8.70%) | 2 (6.45%) |
| 2008 | 3 (13.04%) | 3 (9.68%) |
| 2009 | 1 (4.35%) | 1 (3.23%) |
| 2010 | 4 (17.39%) | 6 (19.35%) |
| 2011 | 4 (17.39%) | 7 (22.58%) |
| 2013 | 1 (4.35%) | 1 (3.23%) |
| Total | 23 (100%) | 31 (100%) |

|  |  |  |
| --- | --- | --- |
| **Year of study publication** | **N per trial (%)** | **N per drug regimen (%)** |
| 2007 | 2 (8.70%) | 2 (6.45%) |
| 2009 | 1 (4.35%) | 1 (3.23%) |
| 2011 | 2 (8.70%) | 2 (6.45%) |
| 2012 | 3 (13.04%) | 4 (12.90%) |
| 2013 | 3 (13.04%) | 3 (9.68%) |
| 2014 | 3 (13.04%) | 5 (16.13%) |
| 2015 | 3 (13.04%) | 5 (16.13%) |
| 2016 | 3 (13.04%) | 4 (12.90%) |
| 2017 | 3 (13.04%) | 5 (16.13%) |
| Total | 23 (100%) | 31 (100%) |

|  |  |  |
| --- | --- | --- |
| **Phase** | **N per trial (%)** | **N per drug regimen (%)** |
| 2 | 17 (73.91%) | 20 (64.52%) |
| 2 | 3 | 1 (4.35%) | 2 (6.45%) |
| 3 | 5 (21.74%) | 9 (29.03%) |
| 4 | 0 (0%) | 0 (0%) |
| Total | 23 (100%) | 31 (100%) |

17 trials were phase 2 (73.91%) and 6 (26.09%) greater than 2. From the 31 drug regimens (one per treatment combination), 20 were of phase 2 (64.52%) and 11 of greater than 2 (35.48%).

|  |  |  |
| --- | --- | --- |
| **Study design** | **N per trial (%)** | **N per drug regimen (%)** |
| randomized | 10 (43.48%) | 18 (58.06%) |
| non-randomized | 13 (56.52%) | 13 (41.94%) |
| Total | 23 (100%) | 31 (100%) |

From the 31 treatment regimens, 18 were randomized (58.06%) and 13 non-randomized (41.94%).

|  |  |  |
| --- | --- | --- |
| **Primary endpoint** | **N per trial (%)** | **N per drug regimen (%)** |
| Progression-Free Survival | 5 (21.74%) | 8 (25.81%) |
| Progression-Free Rate/Survival Rate at 3 months | 5 (21.74%) | 5 (16.13%) |
| Progression-Free Rate/Survival Rate at 6 months | 4 (17.39%) | 5 (16.13%) |
| Clinical Benefit Rate | 3 (13.04%) | 3 (9.68%) |
| Overall Survival | 3 (13.04%) | 6 (19.35%) |
| Response Rate | 3 (13.04%) | 4 (12.90%) |
| Total | 23 (100%) | 31 (100%) |

The most common primary endpoints per trial were Progression-Free Survival and Progression-Free Rate/Survival Rate at 3 months (5 times each).

|  |  |  |
| --- | --- | --- |
| **Estimate used in meta-analysis** | **N per trial (%)** | **N per drug regimen (%)** |
| PFS | 19 (82.61%) | 27 (87.10%) |
| PFR | 4 (17.39%) | 4 (12.90%) |
| Total | 23 (100%) | 31 (100%) |

PFS was used for 19/23 trials and 27/31 rows in our database for all LMS. We used PFR only when information for PFS was not available.

Followingly, we show a table regarding the 31 treatment regimens considered for this review.

|  |  |  |
| --- | --- | --- |
| **Drugs or drug combinations used** | **N (%)** | **Recommended according to ESMO 2018 guidelines?** |
| Doxorubicin | 5 (16.13%) | Yes (for first line) |
| Eribulin | 3 (9.68%) | No (for pre-treated) |
| Docetaxel + Gemcitabine | 2 (6.45%) | No (for first line)  Yes (for pre-treated) |
| Doxorubicin + Evofosfamide | 2 (6.45%) | No (for first line) |
| Pazopanib | 2 (6.45%) | Yes (for pre-treated) |
| Bendamustine | 1 (3.23%) | No (for pre-treated) |
| Brostallicin | 1 (3.23%) | No (for first line) |
| Cixutumumab | 1 (3.23%) | No (for pre-treated) |
| Cyclophosphamide + Sirolimus | 1 (3.23%) | No (for pre-treated) |
| Dacarbazine | 1 (3.23%) | Yes (for pre-treated) |
| Dasatinib | 1 (3.23%) | No (for pre-treated) |
| Doxorubicin + Ifosfamide | 1 (3.23%) | Yes (for first line) |
| Doxorubicin + Trabectedin | 1 (3.23%) | No (for first line) |
| Exatecan | 1 (3.23%) | No (for pre-treated) |
| Gemcitabine | 1 (3.23%) | Yes (for pre-treated) |
| Panobinostat | 1 (3.23%) | No (for pre-treated) |
| Regorafenib | 1 (3.23%) | No (for pre-treated) |
| Ridaforolimus | 1 (3.23%) | No (for pre-treated) |
| Selumetinib | 1 (3.23%) | No (for pre-treated) |
| Selumetinib + Temsirolimus | 1 (3.23%) | No (for pre-treated) |
| Sorafenib | 1 (3.23%) | No (for pre-treated) |
| Trabectedin 3h | 1 (3.23%) | No (for first line) |
| Total | 31 (100%) | //////////////////////////// |

The most common drugs on these trials were Doxorubicin, either alone or in combination, and Eribulin (3 times).

|  |  |  |
| --- | --- | --- |
| **Treatment line** | **N per trial (%)** | **N per drug regimen (%)** |
| first | 7 (30.43%) | 12 (38.71%) |
| pre-treated | 16 (69.57%) | 19 (61.29%) |
| mixed | 0 (0%) | 0 (0%) |
| Total | 23 (100%) | 31 (100%) |

19 excel lines were based on pre-treated population (from 16 studies) and 12 excel lines for first line (7 clinical trials). Trials with mixed lines of treatment were excluded.

|  |  |
| --- | --- |
| **Regimen recommended?** | **N (%)** |
| No | 20 (64.52%) |
| Yes | 11 (35.48%) |
| Total | 31 (100%) |

From the 31 treatment lines selected (23 trials), 20 were non-recommended (64.52%) and 11 were recommended (35.48%).

### **Studies for uterine LMS meta-analyses**

In this subsection descriptive statistics are provided regarding the **9 trials** with uterine LMS included in the meta-analysis, and the corresponding **12 drug regimens** (one line per treatment arm) of the long version of the database.

In the table below, year of study activation (year that first patient was recruited) is presented for the 9 trials in the STBSG database. Most common year of activation at trial level was 2010 (33.33%).

|  |  |  |
| --- | --- | --- |
| **Year of study activation** | **N per trial (%)** | **N per drug regimen (%)** |
| 2002 | 1 (11.11%) | 1 (8.33%) |
| 2003 | 2 (22.22%) | 2 (16.67%) |
| 2006 | 2 (22.22%) | 3 (25.00%) |
| 2009 | 1 (11.11%) | 2 (16.67%) |
| 2010 | 3 (33.33%) | 4 (33.33%) |
| Total | 9 (100%) | 12 (100%) |

Most common year of publication at trial level was 2008 and 2015 (22.22% each).

|  |  |  |
| --- | --- | --- |
| **Year of study publication** | **N per trial (%)** | **N per drug regimen (%)** |
| 2005 | 1 (11.11%) | 1 (8.33%) |
| 2008 | 2 (22.22%) | 2 (16.67%) |
| 2009 | 1 (11.11%) | 1 (8.33%) |
| 2012 | 1 (11.11%) | 2 (16.67%) |
| 2014 | 1 (11.11%) | 1 (8.33%) |
| 2015 | 2 (22.22%) | 3 (25.00%) |
| 2017 | 1 (11.11%) | 2 (16.67%) |
| Total | 9 (100%) | 12 (100%) |

|  |  |  |
| --- | --- | --- |
| **Phase** | **N per trial (%)** | **N per drug regimen (%)** |
| 2 | 7 (77.78%) | 8 (66.67%) |
| 3 | 2 (22.22%) | 4 (33.33%) |
| 4 | 0 (0%) | 0 (0%) |
| Total | 9 (100%) | 12 (100%) |

7 trials were phase 2 (77.78%) and 2 (22.22%) phase 3. From the 12 regimens (one per treatment combination), 8 were of phase 2 (66.67%) and 4 of greater than 2 (33.33%).

|  |  |  |
| --- | --- | --- |
| **Study design** | **N per trial (%)** | **N per drug regimen (%)** |
| randomized | 3 (33.33%) | 6 (50.00%) |
| non-randomized | 6 (66.67%) | 6 (50.00%) |
| Total | 9 (100%) | 12 (100%) |

From the 12 drug regimens analyzed, an equal number came from randomized or non-randomized studies.

|  |  |  |
| --- | --- | --- |
| **Primary endpoint** | **N per trial (%)** | **N per drug regimen (%)** |
| Response Rate | 5 (55.56%) | 6 (50.00%) |
| Progression-Free Survival | 1 (11.11%) | 2 (16.67%) |
| Progression-Free Rate/Survival Rate at 6 months | 1 (11.11%) | 2 (16.67%) |
| Clinical Benefit Rate | 1 (11.11%) | 1 (8.33%) |
| Progression-Free Rate/Survival Rate at 3 months | 0 (0%) | 0 (0%) |
| Overall Survival | 0 (0%) | 0 (0%) |
| Other | 1 (11.11%) | 1 (8.33%) |
| Total | 9 (100%) | 12 (100%) |

The most common primary endpoint per trial and treatment regimen was Response Rate (5 and 6 times).

|  |  |  |
| --- | --- | --- |
| **Estimate used in meta-analysis** | **N per trial (%)** | **N per drug regimen (%)** |
| PFS | 8 (88.89%) | 11 (91.67%) |
| PFR | 1 (11.11%) | 1 (8.33%) |
| Total | 9 (100%) | 12 (100%) |

PFS was used for 8/9 trials and 11/12 rows in our database for uterine LMS. We used PFR only when information for PFS was not available.

Followingly, we show a table regarding the 12 treatment regimens considered for uLMS.

|  |  |  |
| --- | --- | --- |
| **Drugs or drug combinations used** | **N (%)** | **Recommended according to ESMO 2018 guidelines?** |
| Docetaxel + Gemcitabine | 5 (41.67%) | No (for first line)  Yes (for pre-treated) |
| Bevacizumab + Docetaxel + Gemcitabine | 1 (8.33%) | No (for first line) |
| D + M + D + C + S\* | 1 (8.33%) | No (for first line) |
| Doxorubicin | 1 (8.33%) | Yes (for first line) |
| Doxorubicin + Trabectedin | 1 (8.33%) | No (for first line) |
| Gemcitabine | 1 (8.33%) | Yes (for pre-treated) |
| Ixabepilone | 1 (8.33%) | No (for pre-treated) |
| Sunitinib | 1 (8.33%) | No (for pre-treated) |
| Total | 12 (100%) | //////////////////////////// |

\*dacarbazine, mitomycin, doxorubicin, and cisplatin with sargramostim.

The most common drug / drug combination was Docetaxel + Gemcitabine (5 times) in these trials.

|  |  |  |
| --- | --- | --- |
| **Treatment line** | **N per trial (%)** | **N per drug regimen (%)** |
| first | 5 (55.56%) | 7 (58.33%) |
| pre-treated | 4 (44.44%) | 5 (41.67%) |
| mixed | 0 (0%) | 0 (0%) |
| Total | 9 (100%) | 12 (100%) |

7 case were based on first line treatment (from 5 studies) and 5 were based on pre-treated population (4 clinical trials). Trials with mixed lines of treatment were excluded.

|  |  |
| --- | --- |
| **Regimen recommended?** | **N (%)** |
| No | 8 (66.67%) |
| Yes | 4 (33.33%) |
| Total | 12 (100%) |

From the 12 treatment lines selected (9 trials), 4 were recommended (33.33%) and 8 were non-recommended (66.67%).

* 1. **Statistical considerations**

### **Unit of the analyses**

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

### **Meta-analysis methodology**

It was necessary to access for each treatment regimen not only the *effect size* (in terms of PFS at 3/6 months) but also the *sample size* of LMS. To elaborate on this, for Response Rates (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 proportions as the variance of the percentages (calculated under Kaplan-Meier (KM) or binomial distribution) was rarely provided in the papers, the number of cases (pseudocases) was estimated for the LMS sample size and was used 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 (pseudo)cases. For example, assuming that PFS at 3 months is 0.33 with KM methodology for 34 LMS 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 (which is narrower as it takes into account the censoring mechanism).

To quantitatively synthesize the findings of the **23 trials** (7 1L and 16 2L+) for all LMS and **9 trials** (5 1L and 5 2L+) for uLMS, a random-effects model was employed per treatment; suggested by DerSimonian and Laird for meta-analysis of clinical trials.3,4 This model assumes that the true effect of each drug or drug combination is normally distributed. A second assumption made is that effect sizes of drugs are independent, even if they come from the same trial. Here the majority of the trials were single arm (13/23 for all LMS and 6/9 for uLMS). For randomized studies, there might be some dependence as treatment regimens were designed for exactly the same population (also patients were recruited in the same centers).

Outcome variable was the summary rate; the weighted average of the observed effect sizes of individual treatments. Weighting was calculated using the inverse of the total variance (within + between). The main idea of the model was that treatment arms differ in the mixes of participants and in the implementations of the interventions – among other reasons. Therefore, a different effect size was underlying each of them.5 Variation in each treatment arm can be quantified through the true variation in the effect size and the sampling error. Effect sizes (yi) and the sampling variance (vi) were calculated per arm. True effect sizes were not fixed but normally distributed (both within and between regimen 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 drug regimens used in the studies were split into two groups (recommended against non-recommended) using as criterion the 2018 ESMO guidelines6. For both subgroups, the corresponding weight of the effect size of each intervention was estimated using the inverse of the total variance (summary of between and within treatment arm variance). In addition, a mean effect size was estimated for each group together with the 95% CI. Larger arms were given more weight and thus their effect size had greater impact on the overall mean effect.

Moreover, we performed moderator analyses (meta-regressions for drug-level variables) in an attempt to explain a part of between arm variability. Categorical moderators were characteristics which could account for between-arm heterogeneity: dichotomized time period of study activation, phase of the trial, study design and a dichotomized data-driven sample size. These were evaluated first separately in univariate models and significant predictors were added in a multivariate model including the classified drugs (variable of interest) to see if part of the variation is explained with this addition. 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, comparisons were performed directly on the differences in means for the endpoint assuming that there is no difference between the groups. A Z-statistic was computed (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 treatments of the subgroups. Generally, the observed variation in the treatment regimens is attributed to the true variation and to random error. We estimated the -statistic that is insensitive to the number of studies.5 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 treatment arms was performed. In a meta-analysis, there is a small number of studies (treatment arms here) 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 that the effect estimate and its confidence interval do not overlap with the overall group estimate or its CI). We also utilized a series of diagnostics (Baujat plot, influence diagnostics, GOSH plots, normal QQ-plots)7 that have been developed over the years and can substantially contribute in identifying influential studies. The Baujat plot is presented.8 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 treatment regimen to the overall Q-test statistic for heterogeneity versus the influence of each on the summary effect on the y-axis.

As part of this sensitivity analysis, potentially influential treatment arms were removed to check how large their impact was for the pooled estimates and the group difference. Small differences demonstrate findings 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. Bias was examined using common methods for meta-analyses.

Funnel plots were used to investigate risk of publication bias through assessment of the association between treatment regimen effect and size.9,10 X-axis provides the residual values whereas Y-axis a treatment arm’s 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.11,12 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 indicates high risk of publication bias.13 Egger’s regression test is better for smaller number of studies.

### **Proportions and transformation function**

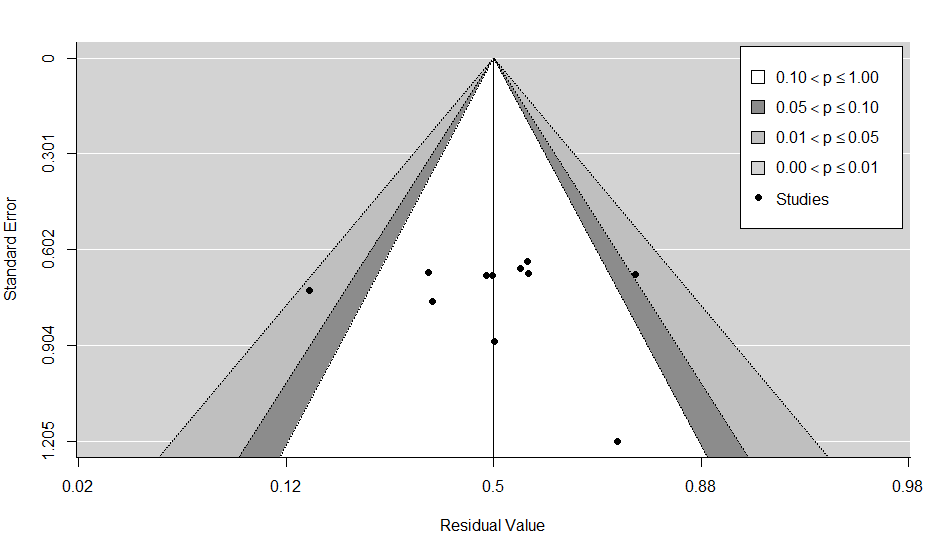
This subsection illustrates the strategy followed regarding the observed proportions of the treatment arms. In case that the observed proportions get further from 0.5 and approach closer to the margins - being smaller than 0.2 or larger than 0.8 – their distribution is increasingly skewed. In such scenario a transformation of the raw proportions is necessary. Appropriate transformation methods include the logit and the Freeman-Tukey double arcsine transformation.14,15 For this project, we applied the logit (log-odds) transformation which is very popular for meta-analyses. The double arcsine transformation could be more appropriate if very small sample sizes/extreme proportions had been observed. Back-transformations of the summary effect were also performed to obtain the effects on the actual scale.

* 1. **Statistical inference for all LMS**

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

The database consisted of 6 trials on first line treatment, corresponding to 11 different therapeutic combinations.

**Risk of bias assessment**

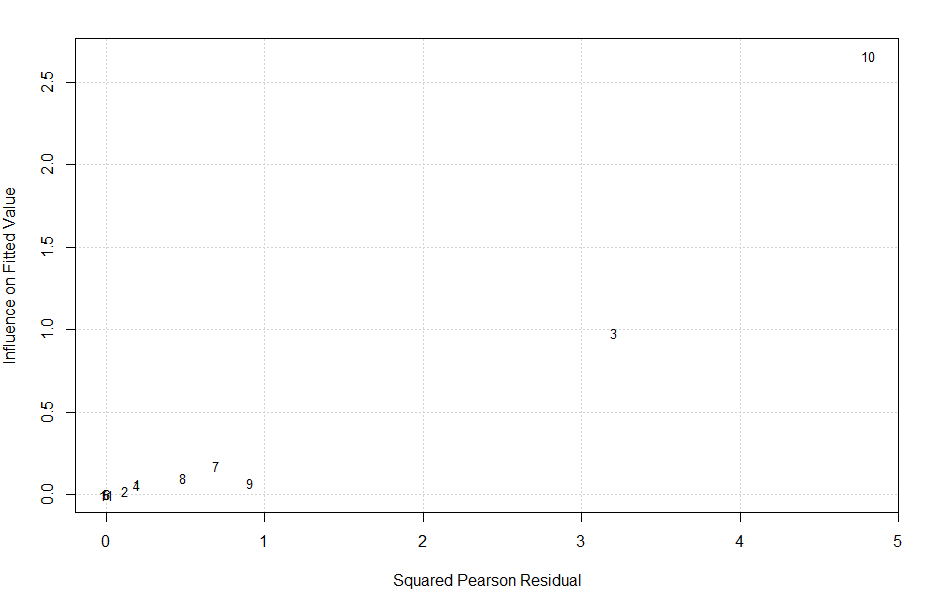
In this subsection, we investigate publication bias based on a funnel plot and 2 formal tests. We show the contour enhanced funnel plot in figure A1. Graph is centered at residual value 0.5 (no effect under the null hypothesis). There are shaped regions for various levels of statistical significance for the points (see the label). Studies are equally spread in the graph. 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 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 only 11 lines in our database, both tests should be interpreted with caution. Rank correlation test provided a p-value of 0.88 and Egger’s regression test also a p-value of 0.88. There is no indication of publication bias from the formal tests.

**Diagnostics for influential studies**

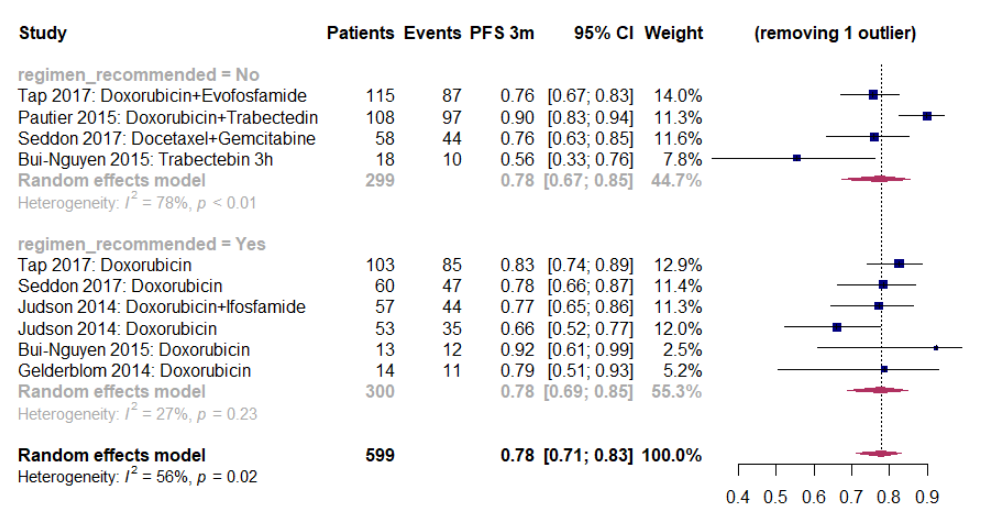
We begin with the examination of the externally standardized residuals of the studies. Z values > 2 or 3 show an influential outlier (its exclusion may lead to significant changes in the fitted meta-analytic model). Study line 10 of Gelderblom 2014: Brostallicin had the highest z value of -3.41 (this study could be influential). We used the Baujat plot (figure A2) as a diagnostic tool to detect sources of heterogeneity and potentially influential studies. Treatment arm with number 10 (Gelderblom 2014:Brostallicin) is projected on the right corner (contributes to heterogeneity and is influential).



**Fig. A2. Baujat plot of the treatment arms for PFS at 3 months (first line patients).**

**Sensitivity meta-analysis**

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



**Fig. A3. Forest plot for PFS rate at 3 months excluding ‘Gelderblom 2014: Brostallicin' (first line patients).**

There is an increase in the overall effect size of the meta-analysis from 0.74 (0.64 – 0.82) to 0.78 (0.71 – 0.83). Overall heterogeneity dropped from 79% (p<0.01) to 56% (p < 0.02). This treatment arm was removed from the non-recommended treatment regimens. The estimate of the subgroup has been recalculated to 0.78 (0.67 – 0.85) from 0.69 (0.53 – 0.82). There is a small decrease in the group heterogeneity to 78% (from 90%). Findings indicate that ‘Gelderblom 2014: Brostallicin' is an influential study. Again, the difference between a regimen recommended versus a non-recommended is not statistically significant based on moderator’s test (p = 0.97). Note that the combination of Doxorubicin + Trabectedin in the Pautier study had a 0.90 PFS proportion at 3 months, which shows that it should have been investigated further (even if it is not recommended in ESMO 2018 guidelines). There is an ongoing phase III study investigating this therapeutic combination: <https://clinicaltrials.gov/ct2/show/NCT02997358>.

**Meta-regressions**

The effect of potential moderators was examined on the effect size (rate) through meta-regressions. First, moderators were evaluated separately in univariate models. If any moderator was found significant, we then test it simultaneously in a multivariate model including the regimen recommended (variable of main interest).

The moderators used for these meta-regressions are presented below. For year of activation and sample size, we performed the dichotomization based on the median values of the entire database for “all” Leiomyosarcomas. We only performed analyses meta-regression for moderators with more than two studies.

|  |  |
| --- | --- |
| **Moderator** | **Dichotomization** (first group reference, count of treatment regimens in parenthesis) |
| Phase of the trial | 2 (3), greater than 2 (8) |
| Study design | Randomized (10), non-randomized (1) |
| Year of activation | 2000-2008 (7), 2009-2013 (4) |
| Sample size | 10-38 (4), >38 (7) |

Here there was only one non-randomized study so this moderator was not considered. 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 reference level)** | **Test for moderator**  **P-value** | **Residual heterogeneity (I2)** | **Test for residual heterogeneity**  **P-value** |
| Phase of the trial:  greater than 2 | 0.76 (vs 0.70) | 0.61 | 80.89% | < 0.01 |
| Year of activation:  2009-2013 | 0.63 (vs 0.79) | 0.06 | 74.22% | < 0.01 |
| Sample size: >38 | 0.79 (vs 0.57) | 0.02 | 69.47% | < 0.01 |

For the phase of the trial, residual heterogeneity was high (I2 = 80.89%, p < 0.01). 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 0.61 (not statistically significant).

For the year of activation, residual heterogeneity was high (I2 = 74.22%, p < 0.01). A year of activation between 2009 and 2013 had a negative effect on the effect size, but not significant against 2000-2008 (p = 0.06).

For sample size, residual heterogeneity was again high (I2 = 69.47%, p < 0.01). A sample size above 38 patients has a positive effect on the observed rate versus ≤ 38 which is significant (p = 0.02).

We also performed a *multivariate* meta-regression adjusting for whether a regimen was recommended with sample size:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Moderator (multivariate analysis)** | **Adjusted estimated PFS (vs reference level)** | **Z-statistic**  **P-value** | **Residual heterogeneity (I2)** | **Test for residual heterogeneity**  **P-value** |
| Regimen:  Recommended | 0.62 (vs 0.56\*) | 0.50 | 72.82% | < 0.01 |
| Sample size: >38 | 0.76 (vs 0.56\*) | 0.05 | 72.82% | < 0.01 |

\*Regimen not recommended and sample size 10-38 patients.

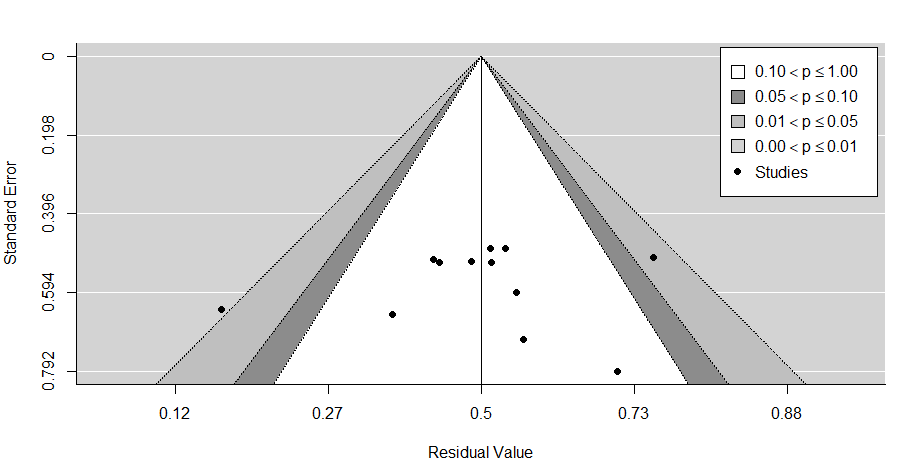
Regimen recommended versus non-recommended has a positive effect (p-value = 0.50, not significant) and sample size a bit smaller positive effect compared to the unadjusted (still statistically significant p-value = 0.05). The omnibus chi-square test for both moderators is not significant (p = 0.09). Residual heterogeneity remains very high (I2 = 72.82%, p < 0.01 vs I2 = 80.50%, p < 0.01 for the model including the classified drugs only). This shows that a sample size > 38 cannot explain a substantial part of the model’s residual heterogeneity.

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

The database for this analysis consisted of 7 trials covering 12 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. A4 detected one study in the grey zone of the right side (0.01 ≤ p ≤ 0.05) and another study on the light grey zone of the left side (p < 0.01). 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 0.5 and 0.64, respectively. There is no indication of publication bias.

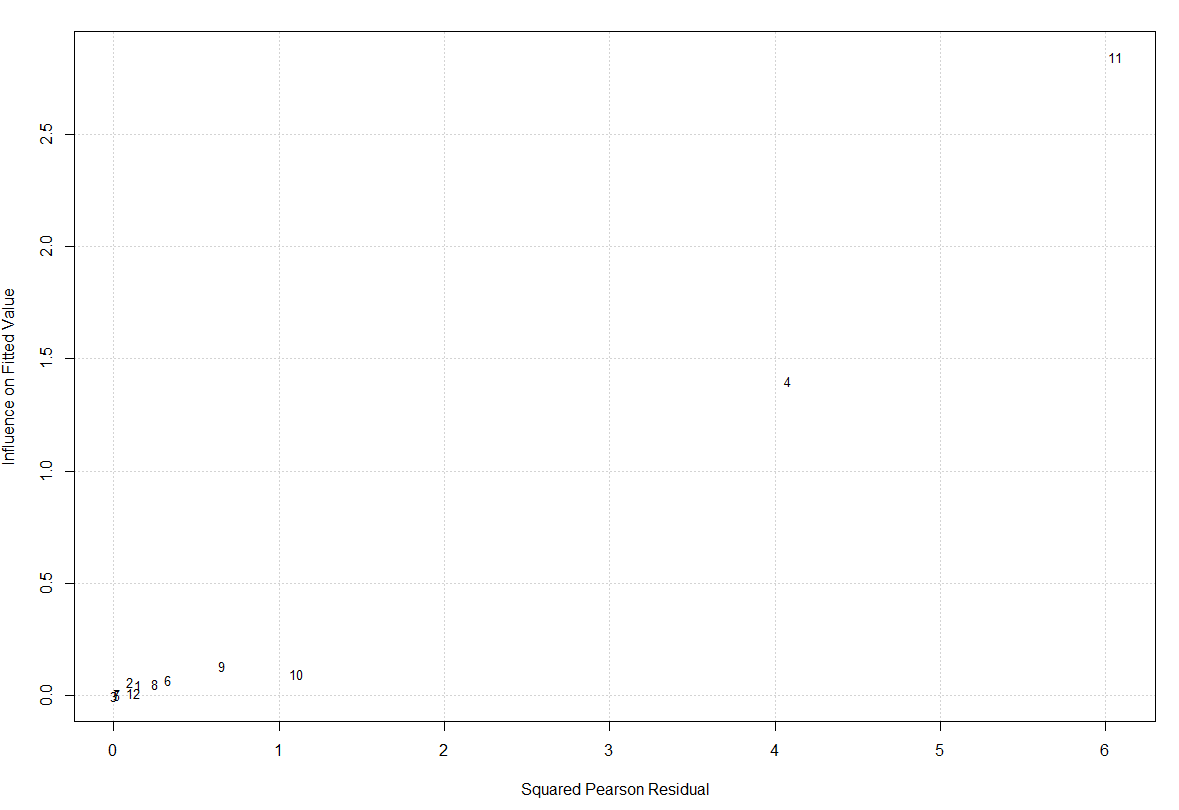


**Fig. A4. Contour enhanced funnel plot for PFS at 6 months (first line patients).**

**Diagnostics for influential studies**

Externally standardized residuals produced the highest z-value of -3.33 for the study of Gelderblom 2014: Brostallicin. This suggests that it might be an influential outlier.

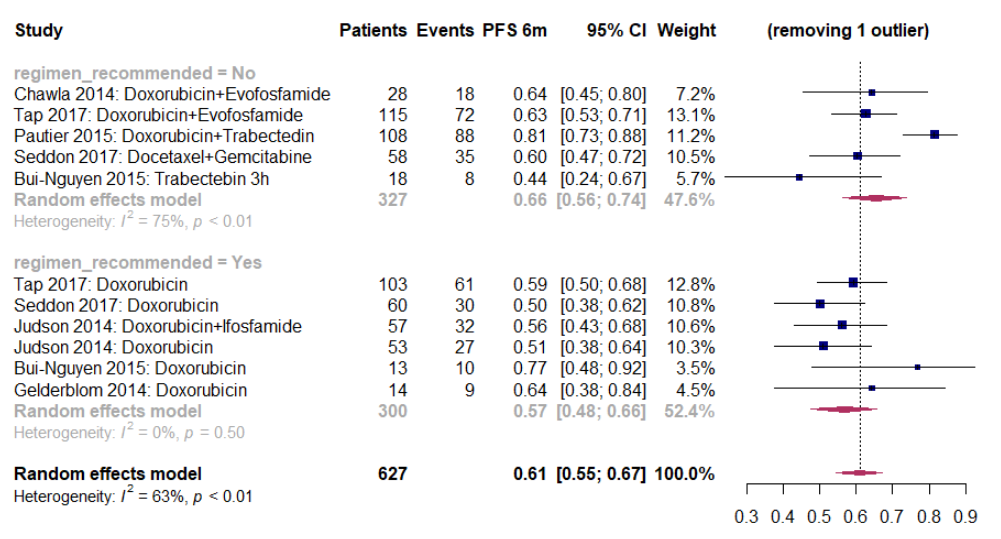
The Baujat plot in fig. A5 projects line with number 11 (Gelderblom 2014) on the right corner. This study contributes to heterogeneity and is the most influential for the meta-analysis.



**Fig. A5. Baujat plot of the treatment arms for PFS at 6 months (first line patients).**

**Sensitivity meta-analysis**

A sensitivity analysis was performed removing Gelderblom 2014:Brostallicin from the database and re-fitting a random effects model.



**Fig. A6. Forest plot for PFS at 6 months excluding ‘Gelderblom 2014: Brostallicin' (first line patients).**

Overall effect size of the meta-analysis increased to 0.61 (0.55-0.67) from 0.58 (0.50-0.66). Overall heterogeneity dropped from 74% (p < 0.01) to 63% (p < 0.01). The drug regimen was removed from the non-recommended treatments. The estimate of the subgroup has been recalculated to 0.66 (0.56 – 0.74) from 0.59 (0.47 – 0.70). Heterogeneity was decreased from 85% (p < 0.01) to 75% (p < 0.01). Test of moderator with drugs grouped as recommended or not was not significant (p = 0.18).

**Meta-regressions**

|  |  |
| --- | --- |
| **Moderator** | **Dichotomization** (first group reference, count of treatment regimens in parenthesis) |
| Phase of the trial | 2 (4), greater than 2 (8) |
| Study design | Randomized (10), non-randomized (2) |
| Year of activation | 2000-2008 (8), 2009-2013 (4) |
| Sample size | 10-38 (5), >38 (7) |

Again, there were only two non-randomized studies so study design was not considered as a moderator.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Moderator**  **(univariate analysis)** | **Estimated PFS (vs reference level)** | **Test for moderator**  **P-value** | **Residual heterogeneity (I2)** | **Test for residual heterogeneity**  **P-value** |
| Phase of the trial:  greater than 2 | 0.57 (vs 0.61) | 0.65 | 72.84% | < 0.01 |
| Year of activation:  2009-2013 | 0.48 (vs 0.63) | 0.08 | 70.51% | < 0.01 |
| Sample size: >38 | 0.61 (vs 0.52) | 0.34 | 74.30% | < 0.01 |

For treatments of phase greater than 2 the effect size is smaller. The omnibus test for the effect of the moderator provided a p-value of 0.65, not statistically significant. Residual heterogeneity was high (I2 = 72.84%, p < 0.01).

A year of activation between 2009 and 2013 had a negative effect on the PFS against a study activated between 2000 - 2008, but not statistically significant (p = 0.08). Residual heterogeneity was high (I2 = 70.51%, p < 0.01).

A sample size above 38 patients had a positive effect on the observed proportion, however not significant (p = 0.34). Again, residual heterogeneity is high (I2 = 74.30%, p < 0.01).

No multivariate moderator analysis was performed as none of these moderators was statistically significant.

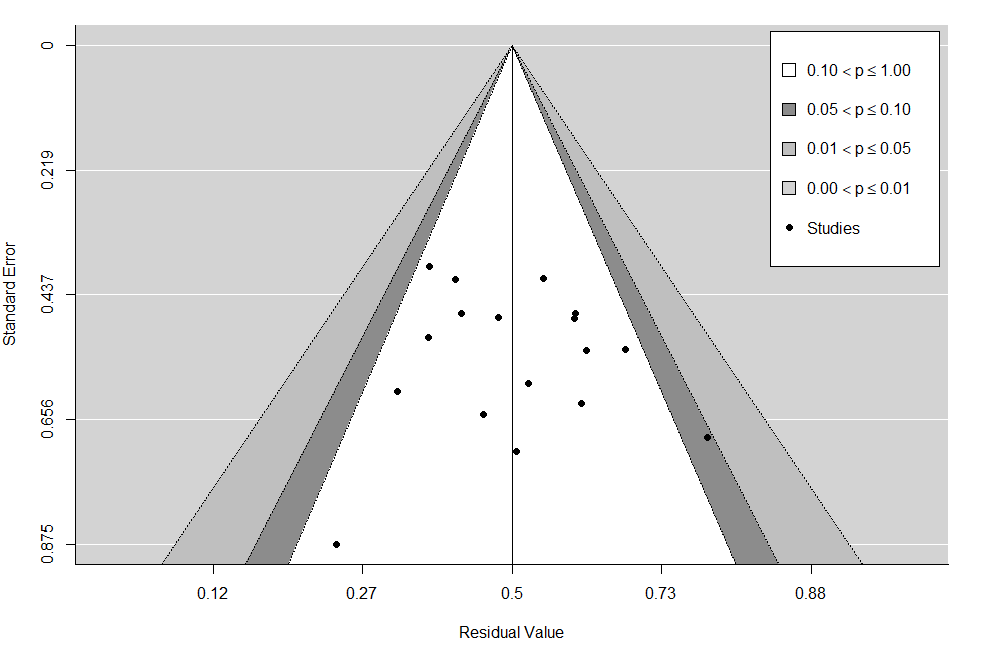
### **Analysis for pre-treated patients at 3 months**

The database consisted of 14 trials, corresponding to 17 different therapeutic combinations.

**Risk of bias assessment**

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

In figure A7, there is one study in the dark grey zone (right side, 0.05 < p ≤ 0.10). Studies are symmetrically distributed. The rank correlation test produced a p-value of 0.97 and the Egger’s regression test a p-value of 0.44. There is no indication of publication from the formal examination diagnostics.

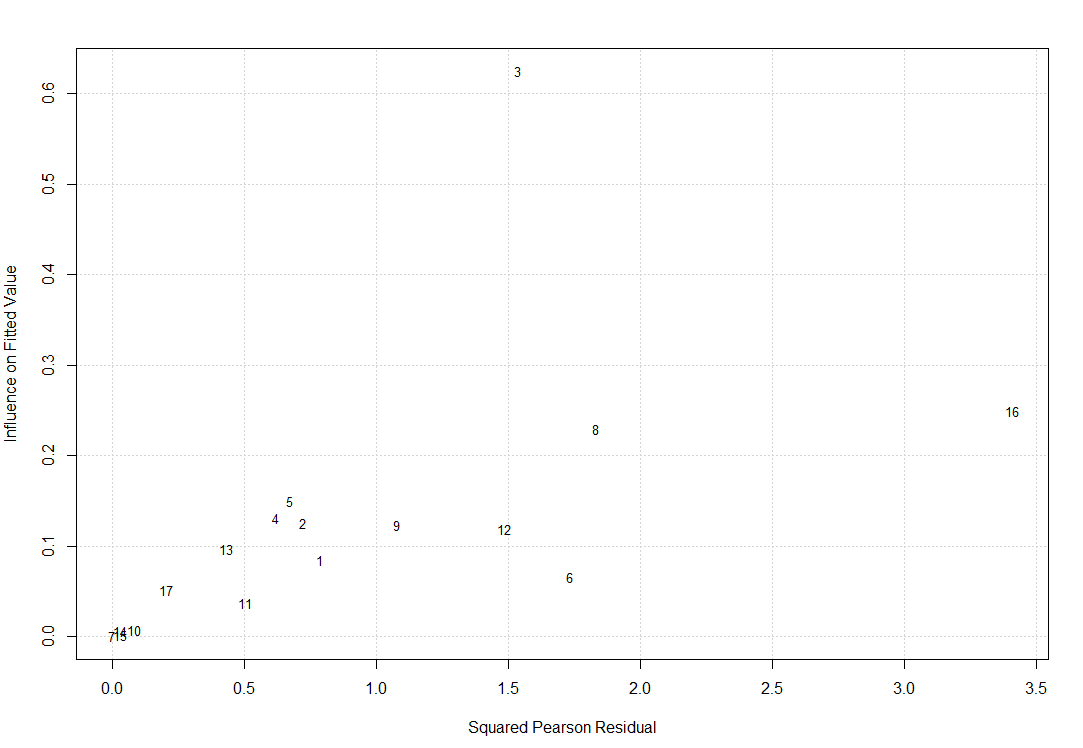


**Fig. A7. Contour enhanced funnel plot for PFS at 3 months (pre-treated patients).**

**Diagnostics for influential studies**

Examination of the externally standardized residuals showed that study in line 16 ‘Schuetze 2012: Cyclophosphamide+Sirolimus’ had the highest z-value of 1.98 so it was potentially influential for the meta-analysis.

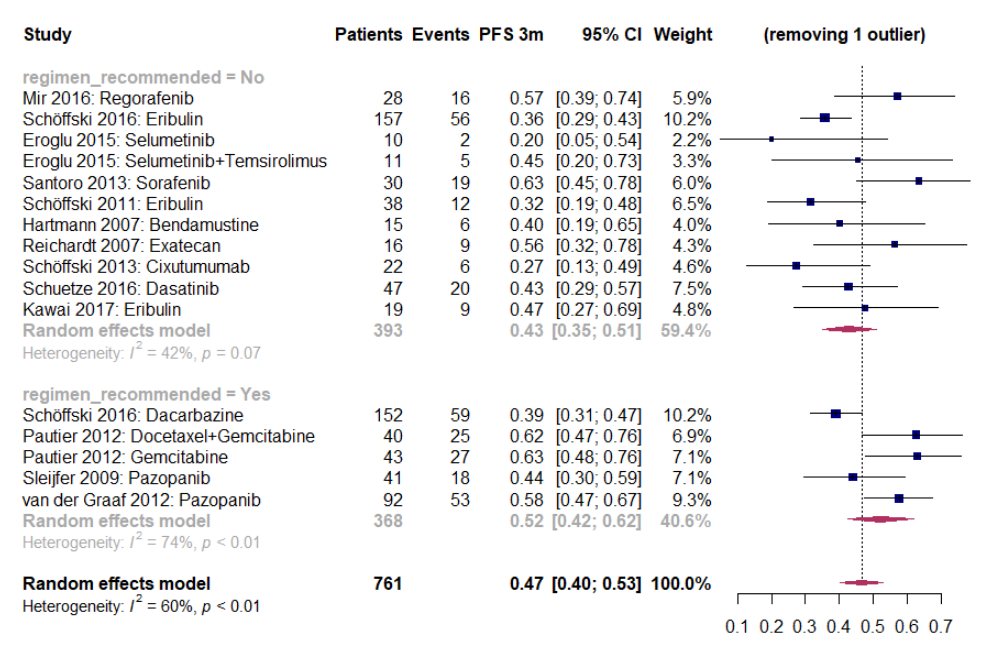
The Baujat plot in fig. A8 projected line with number 16 on the upper right corner. ‘Schuetze 2012: Cyclophosphamide+Sirolimus’ seemed to contribute to the highest amount of heterogeneity (based on squared Pearson residuals) and was the second most influential (fitted values).



**Fig. A8. Baujat plot of the treatment arms for PFS at 3 months (pre-treated patients).**

**Sensitivity meta-analysis**

A sensitivity meta-analysis was performed to focus on the impact of ‘Schuetze 2012: Cyclophosphamide+Sirolimus’ for the random-effects model.



**Fig. A9. Forest plot for PFS at 3 months excluding ‘Schuetze 2012: Cyclophosphamide+Sirolimus’ (pre-treated patients).**

There is a very slight decrease in the overall effect size from 0.48 (0.41 – 0.54) to 0.47 (0.40 – 0.53). Overall heterogeneity was reduced to 60% p < 0.01 (from 63% p < 0.01). Study was removed from the non-recommended group of drugs. The recalculated estimate for this group was 0.43 (0.35 – 0.51) from 0.45 (0.37 – 0.53). Heterogeneity of the non-recommended subgroup dropped from 53% (p = 0.02) to 42% (p = 0.07). Based on the omnibus chi-square test for moderators, there was no difference between the subgroups (p = 0.14). Findings indicated that the excluded study was not that much influential.

**Meta-regressions**

|  |  |
| --- | --- |
| **Moderator** | **Dichotomization** (first group reference, count of treatment regimens in parenthesis) |
| Phase of the trial | 2 (14), greater than 2 (3) |
| Study design | Randomized (8), non-randomized (9) |
| Year of activation | 2000-2008 (6), 2009-2013 (11) |
| Sample size | 10-38 (10), >38 (7) |

Potential moderators were examined for the PFS rate at 3 months (pre-treated patients).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Moderator**  **(univariate analysis)** | **Estimated PFS (vs reference level)** | **Test for moderator**  **P-value** | **Residual heterogeneity (I2)** | **Test for residual heterogeneity**  **P-value** |
| Phase of the trial:  greater than 2 | 0.44 (vs 0.49) | 0.45 | 60.90% | < 0.01 |
| Study design:  non-randomized | 0.46 (vs 0.49) | 0.71 | 64.94% | < 0.01 |
| Year of activation:  2009-2013 | 0.51 (vs 0.41) | 0.10 | 51.09% | 0.01 |
| Sample size: >38 | 0.48 (vs 0.47) | 0.86 | 64.75% | < 0.01 |

Studies of phase greater than 2 decreased the effect size. The omnibus test for the effect of the moderator based in a chi-square distribution with 1 df provided a p-value of 0.45 (not significant). Residual heterogeneity was high (I2 = 60.90%, p < 0.01). This moderator was not able to explain part of the heterogeneity.

For the pre-treated population, a non-randomized design decreased the effect size at 3 months. There was no difference between randomized and non-randomized trials (p = 0.71). Residual heterogeneity was high (64.94%, p < 0.01).

A year of activation between 2009-2013 had a positive effect on the estimate. However, there no statistically significant difference between the 2 subgroups (2009-2013 versus 2000-2008) as p = 0.10. Residual heterogeneity is moderate to high (51.09%, p = 0.01).

A sample size above 38 patients had a slightly positive effect on the observed rate versus a sample size of 38 or less. This effect was not statistically significant (p = 0.86). Residual heterogeneity is high (I2 = 64.75%, p < 0.01).

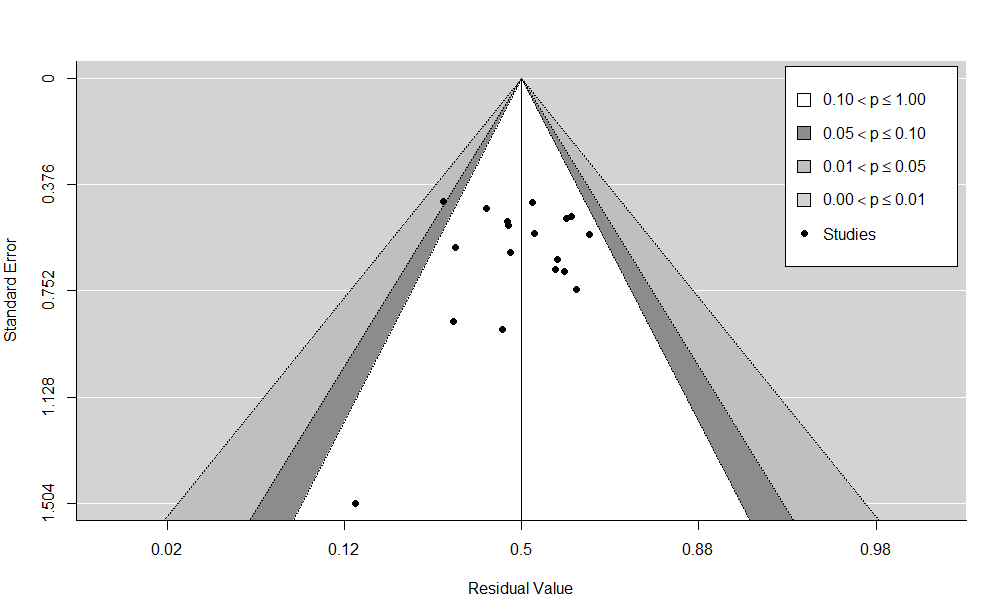
No moderators were found significant. Consequently, no multivariate moderator analysis was performed.

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

For this case, the EORTC database consisted of 15 trials and 18 treatment monotherapies or combinations.

**Risk of bias assessment**

The contour enhanced funnel plot showed 1 study with 0.01 ≤ p ≤ 0.05. There is no indication of publication bias based on this plot. There is symmetry apart from 3 studies on the lower part of the left side.



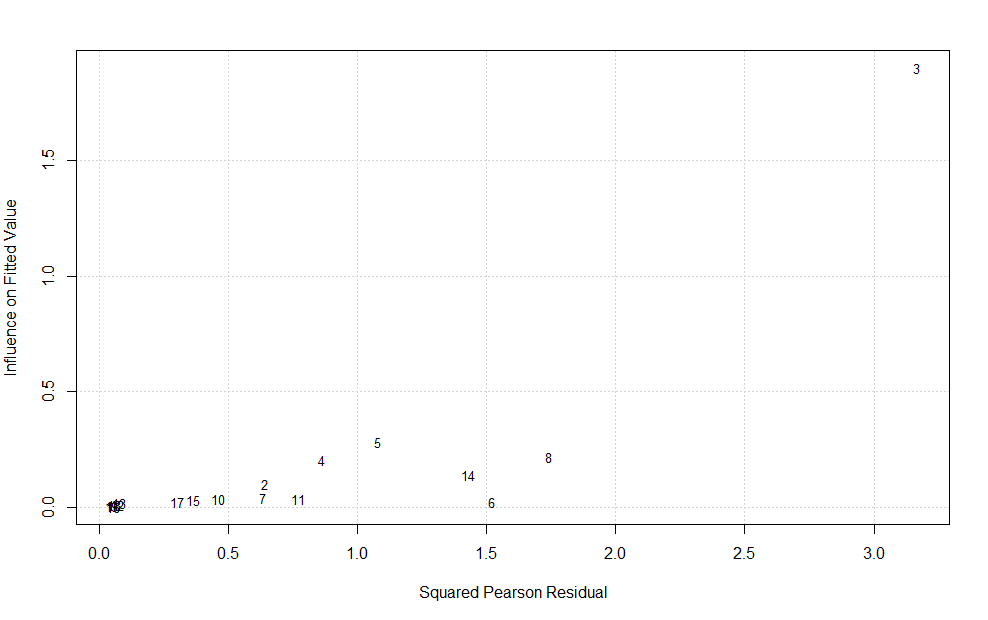
**Fig. A10. Contour enhanced funnel plot for PFS at 6 months (pre-treated patients).**

We also performed the 2 formal tests. The rank correlation test produced a p-value of 0.60 and the Egger’s regression test a p-value of 0.80. There is no indication of publication bias from the formal tests.

**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 2 with study ‘Schöffski 2016: Dacarbazine’ had the highest z-value of -3.15.

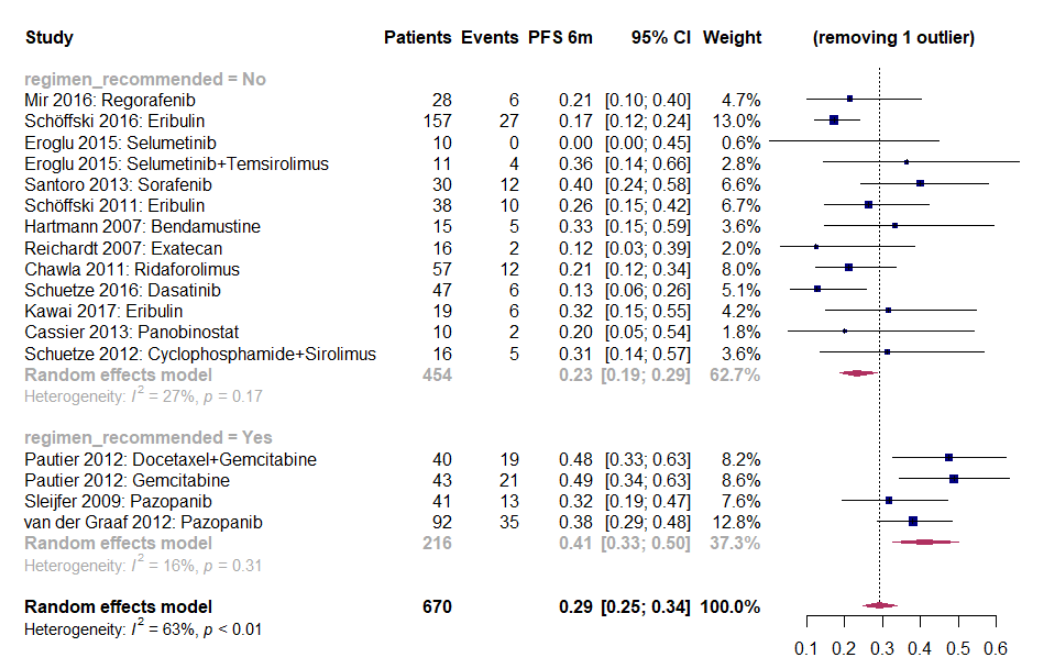
The baujat plot (fig. A11) projected this study to the upper right corner. This study contributes the most to heterogeneity (squared Pearson residuals) and it is the most influential based on the fitted values on the y-axis.



**Fig. A11. Baujat plot of the treatment arms 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. A12). The overall effect size slightly increased to 0.29 (0.25 – 0.34) from 0.28 (0.22 – 0.34). Overall heterogeneity dropped from 66% to 63% (p < 0.01). Study was removed from the recommended group of drugs. The estimated effect size of this subgroup increased to 0.41 (0.33 – 0.50) from 0.35 (0.26 – 0.46). Heterogeneity of the subgroup decreased substantially from 83% (p < 0.01) to 16% (p = 0.31). 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 6 ‘Eroglu 2015: Selumetinib’ had 0/10 patients progression-free and alive at 6 months but it was not influential because of its small weight (0.6%).



**Fig. A12. Forest plot for PFS at 6 months excluding ‘Schöffski 2016: Dacarbazine' (pre-treated patients).**

**Meta-regressions**

|  |  |
| --- | --- |
| **Moderator** | **Dichotomization** (first group reference, count of treatment regimens in parenthesis) |
| Phase of the trial | 2 (15), greater than 2 (3) |
| Study design | Randomized (8), non-randomized (10) |
| Year of activation | 2000-2008 (7), 2009-2013 (11) |
| Sample size | 10-38 (10), >38 (8) |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Moderator**  **(univariate analysis)** | **Estimated PFS (vs reference level)** | **Test for moderator**  **P-value** | **Residual heterogeneity (I2)** | **Test for residual heterogeneity**  **P-value** |
| Phase of the trial:  greater than 2 | 0.23 (vs 0.29) | 0.40 | 63.86% | < 0.01 |
| Study design:  non-randomized | 0.26 (vs 0.30) | 0.57 | 67.89% | < 0.01 |
| Year of activation:  2009-2013 | 0.32 (vs 0.21) | 0.03 | 47.28% | 0.02 |
| Sample size: >38 | 0.28 (vs 0.27) | 0.94 | 67.91% | < 0.01 |

Studies of phase greater than 2 decreased 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 0.40 (not significant). Residual heterogeneity was high (I2 = 63.86%, p < 0.01).

A non-randomized study versus randomized decreased the effect size. The difference between the groups was statistically not significant (p-value = 0.57). Residual heterogeneity was high (I2 = 67.89%, p < 0.01).

A study activated between 2009 – 2013 versus 2000 – 2008 increased the effect size. This increase was significant (p = 0.03). Year of activation could explain part of the heterogeneity as I2  reduced (I2 = 47.28%, p = 0.02).

A sample size > 38 versus 10-38 increased slightly the estimate. Based on the test for moderators this difference was not significant (p = 0.94). Residual heterogeneity was high (I2 = 67.91%, p < 0.01).

Here, we performed a multivariate meta-regression adjusting regimen recommended with year of activation of the trial:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Moderator (multivariate analysis)** | **Adjusted estimated PFS (vs reference level)** | **Z-statistic**  **P-value** | **Residual heterogeneity (I2)** | **Test for residual heterogeneity**  **P-value** |
| Regimen:  Recommended | 0.26 (vs 0.18\*) | 0.06 | 38.82% | 0.06 |
| Year of activation:  2009-2013 | 0.28 (vs 0.18\*) | 0.04 | 38.82% | 0.06 |

\*Regimen not recommended and year of activation 2000-2008.

As a final meta-regression, we adjusted regimen recommended with the year of activation (significant moderator based on the univariate analyses). The test of moderators examining the two variables together versus the null model was significant (p < 0.01). The multivariate meta-regression model was able to explain a larger part of the residual heterogeneity (I2 = 38.82%, p = 0.06) compared with the univariate model for drug regimen (I2 = 59.91%, p < 0.01). This indicates that year of activation can explain some part of the residual heterogeneity of the random effects model (remaining variability between the treatment regimens not accounted for).

* 1. **Statistical inference for uterine LMS**

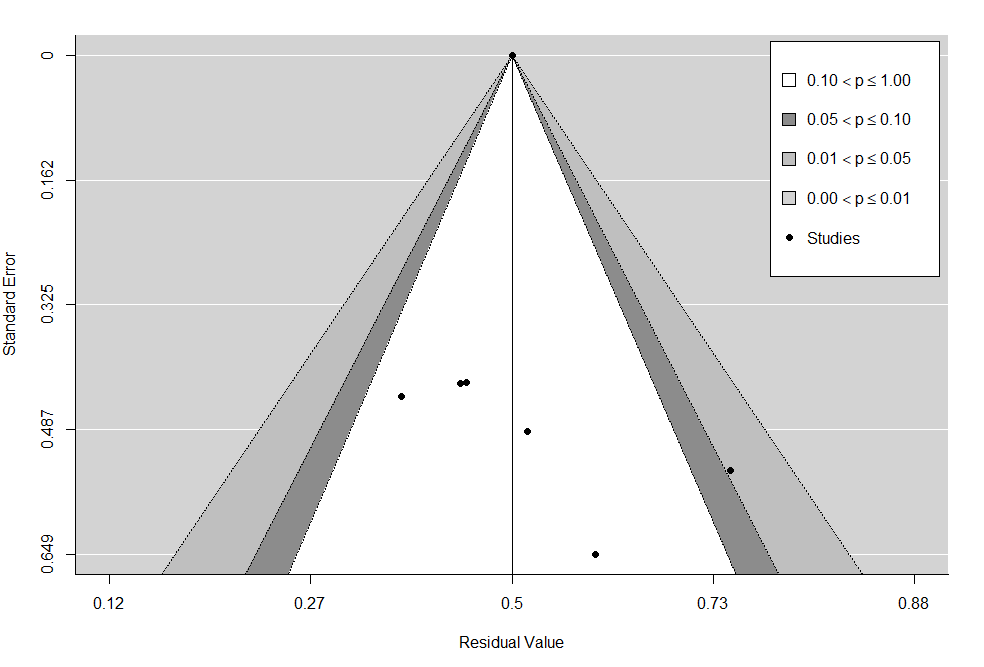
Meta-regressions were not performed for uterine LMS due to the limited number of treatment arms for first line or pre-treated patients.

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

The database consisted of 5 trials on first line treatment, corresponding to 7 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 13. 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 was only one study at the bottom right side of the graph. Studies are more or less equally spread for lower standard error. One study was in the grey zone (0.01 ≤ p ≤ 0.05) on the right side. There might be some publication bias here.



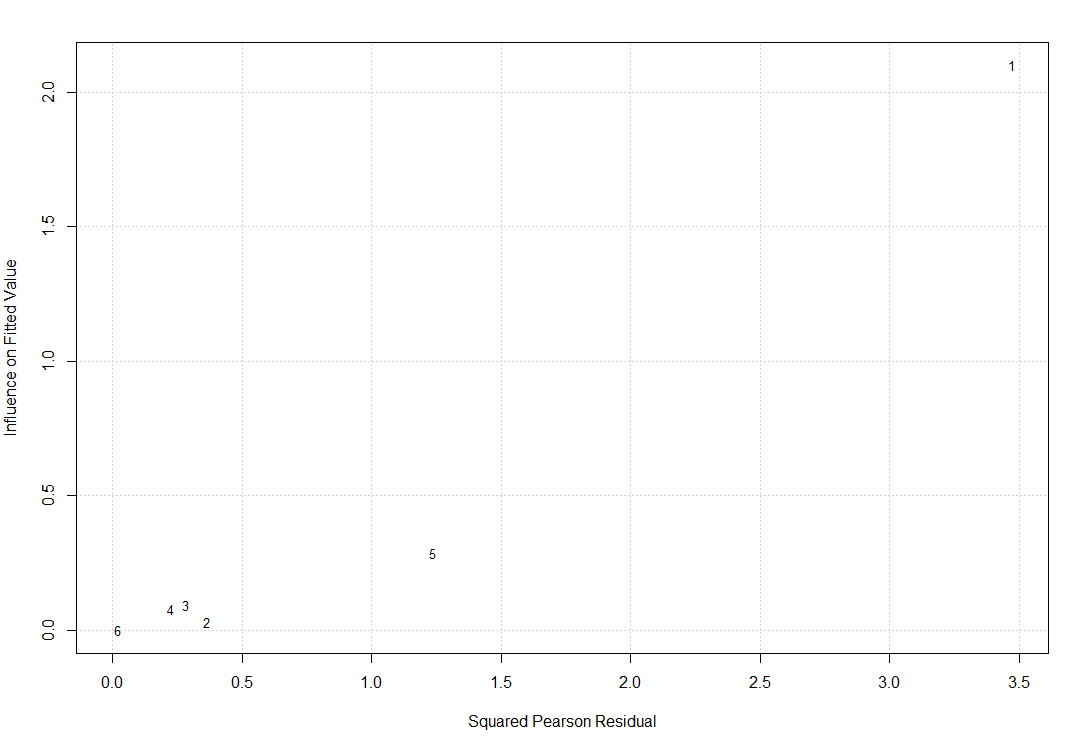
**Fig. A13. Contour enhanced funnel plot for PFS at 3 months (first line patients).**

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 0.14 and Egger’s regression test a p-value of 0.03. There is high risk of publication bias according to Egger’s regression test (better for smaller number of studies). It is possible that the rank correlation test did not detect publication bias, being underpowered due to the limited number of rows.

**Diagnostics for influential studies**

We examined 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 1 of Pautier 2015: Doxorubicin+Trabectedin had a value of 2.81.

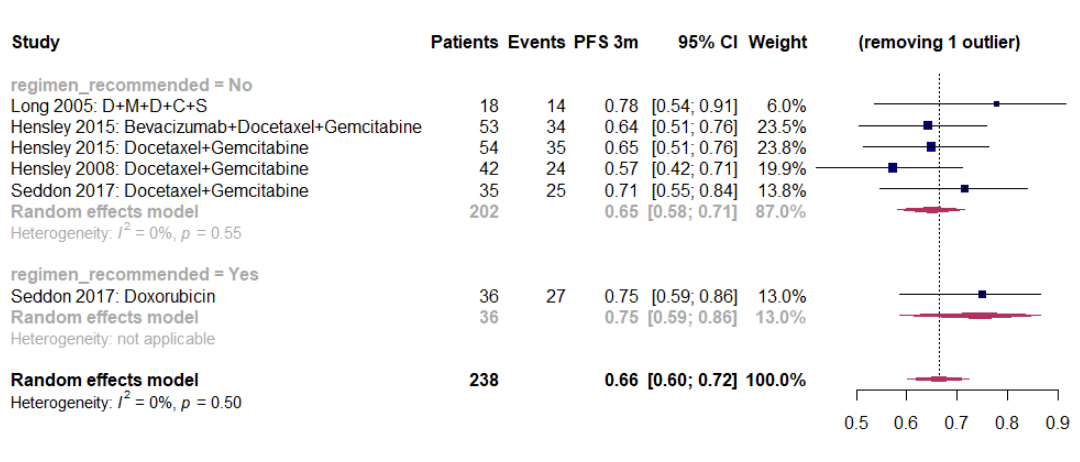
We used the Baujat plot (figure A14) as a diagnostic tool to detect sources of heterogeneity and potentially influential studies. Study with number 1 (Pautier 2015: Doxorubicin+Trabectedin) was projected on the right upper corner (contributes the most to heterogeneity and is the most influential).



**Fig. A14. Baujat plot of the treatment arms for PFS at 3 months (first line patients).**

**Sensitivity meta-analysis**

To investigate the effect of ‘Pautier 2015: Doxorubicin+Trabectedin’ which was found as influential in previous section, a sensitivity analysis was performed removing this treatment arm from the analysis and re-fitting the random-effects model.



**Fig. A15. Forest plot for PFS at 3 months excluding ‘Pautier 2015: Doxorubicin+Trabectedin' (first line patients). D+M+D+C+S = dacarbazine, mitomycin, doxorubicin, and cisplatin with sargramostim.**

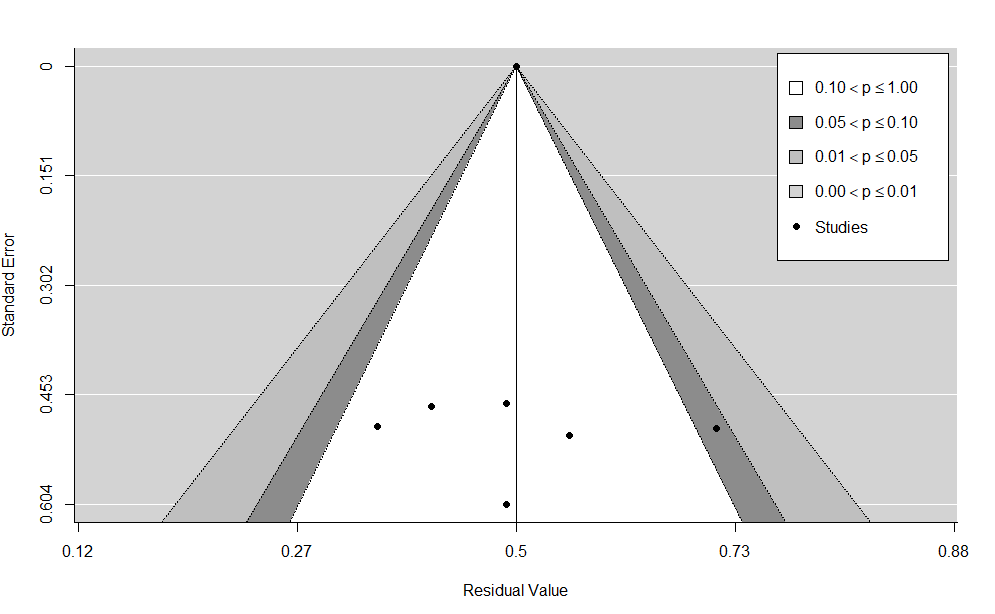
There is a decrease in the overall effect size of the meta-analysis from 0.71 (0.62 – 0.78) to 0.66 (0.60 – 0.72). Overall heterogeneity dropped significantly from 48% (p=0.07) to 0% (p = 0.50). This treatment arm was removed from the non-recommended treatment regimens. The estimate of the subgroup has been recalculated to 0.65 (0.58 – 0.71) from 0.70 (0.60 – 0.78). There is a big decrease in the group heterogeneity to 0%, p = 0.55 (from 54%, p = 0.05). Findings indicate that ‘Pautier 2015: Doxorubicin+Trabectedin' is an influential study. Again, the difference between a regimen recommended versus a non-recommended was not statistically significant based on moderator’s test (p = 0.25). This sensitivity meta-analysis indicates that the results of the 3-month PFS for the uLMS first line patients are not that much robust which was expected given the limited number of available studies.

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

The database consisted of 5 trials covering 7 different therapeutic combinations.

**Risk of bias assessment**

Again, we investigated publication bias with a funnel plot and 2 formal tests. The contour enhanced funnel plot in fig. A16 detected one study in the dark grey zone of the right side (0.05 ≤ p ≤ 0.10). There is only one study at the bottom left side of the graph and the rest of the studies are more or less symmetrically spread. Publication bias cannot be excluded.



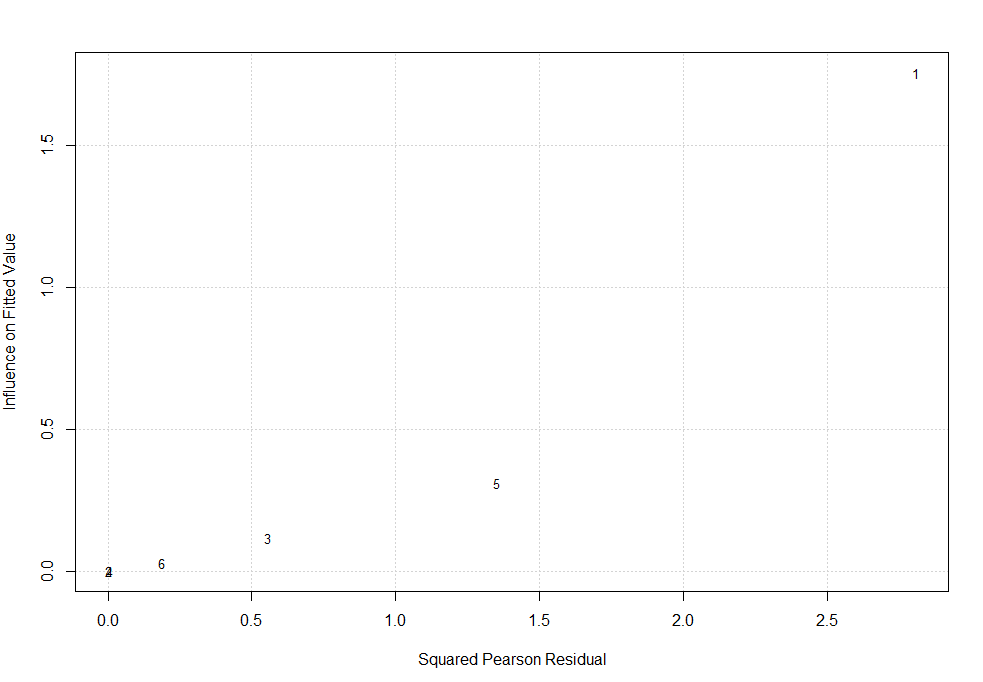
**Fig. A16. Contour enhanced funnel plot for PFS at 6 months (first line patients).**

Hence, we examined the risk of bias more formally. The p-values from the rank correlation and the Egger’s regression tests were 1.00 and 0.78, respectively. There is no indication of publication bias here. Note that these results could have been affected by the limited number of studies.

**Diagnostics for influential studies**

Examination of the externally standardized residuals showed that study in line 1 ‘Pautier 2015: Doxorubicin + Trabectedin’ had the highest z-value of 3.00, so it was potentially influential for the meta-analysis.

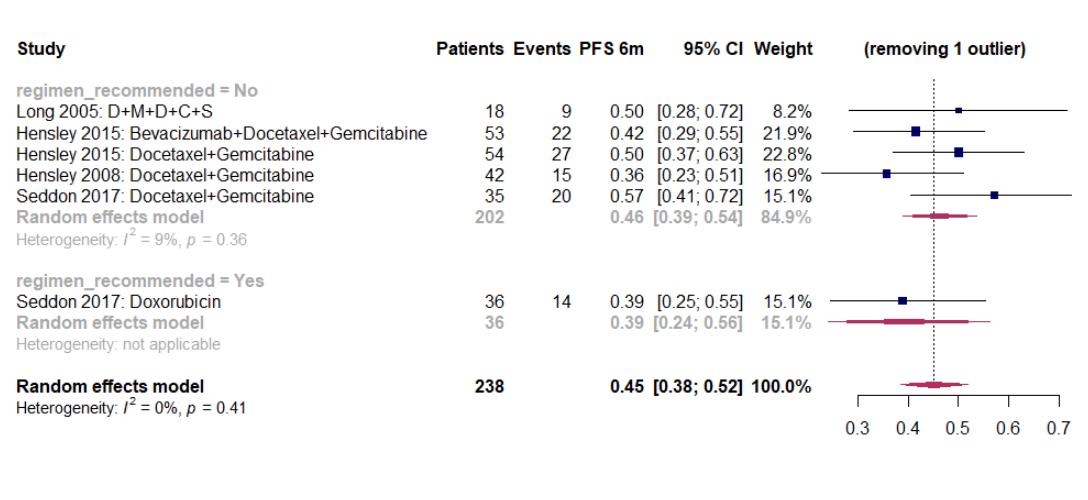
The Baujat plot in fig. A17 projected line with number 1 on the upper right corner. ‘Pautier 2015: Doxorubicin + Trabectedin’ contributed to the highest amount of heterogeneity (based on squared Pearson residuals) and was the most influential (fitted values).



**Fig. A17. Baujat plot of the treatment arms for PFS at 6 months (first line patients).**

**Sensitivity meta-analysis**

A sensitivity analysis was performed removing the study of Pautier 2015: Doxorubicin + Trabectedin from the database and re-fitting a random effects model.



**Fig. A18. Forest plot for PFS at 6 months excluding ‘Pautier 2015: Doxorubicin + Trabectedin ' (first line patients). D+M+D+C+S = dacarbazine, mitomycin, doxorubicin, and cisplatin with sargramostim.**

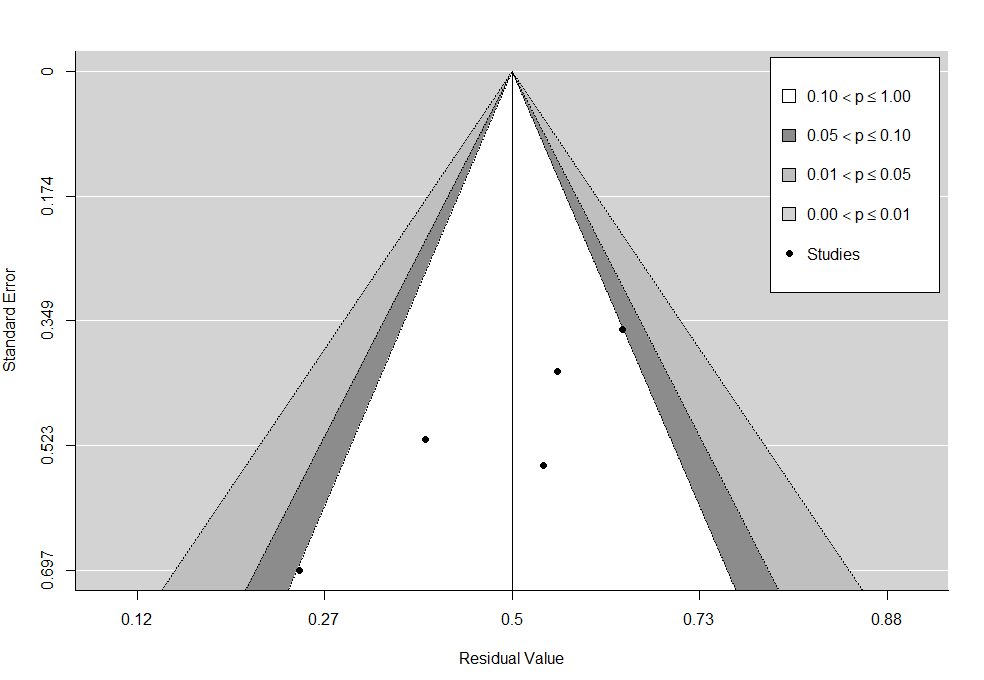
Overall effect size of the meta-analysis decreased to 0.45 (0.38-0.52) from 0.49 (0.39 – 0.59). Overall heterogeneity dropped a lot from 62% (p = 0.01) to 0% (p = 0.41). The study was removed from the non-recommended treatments. The estimate of the subgroup has been recalculated to 0.46 (0.39 – 0.54) from 0.51 (0.40 – 0.62). Heterogeneity decreased significantly from 65% (p=0.01) to 9% (p = 0.36). Test of moderator with drugs grouped as recommended or not was not significant (p = 0.44). Again, results were not very robust which can be attributed to the quite small number of studies (7 rows in long database).

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

The database consisted of just 4 trials covering 5 different therapeutic combinations.

**Risk of bias assessment**

We examined the risk of publication bias with a contour-enhanced funnel plot and formal diagnostic tests. In figure A19, there were two studies in the dark grey zone (left and right side, 0.05 < p ≤ 0.10) as well as asymmetry in the distribution of the studies which signifies potential risk of bias. The rank correlation test produced a p-value of 0.48 and the Egger’s regression test a p-value of 0.05 (better than rank correlation test for smaller number of studies). There was an indication of high risk for publication bias from the formal examination diagnostic of Egger.

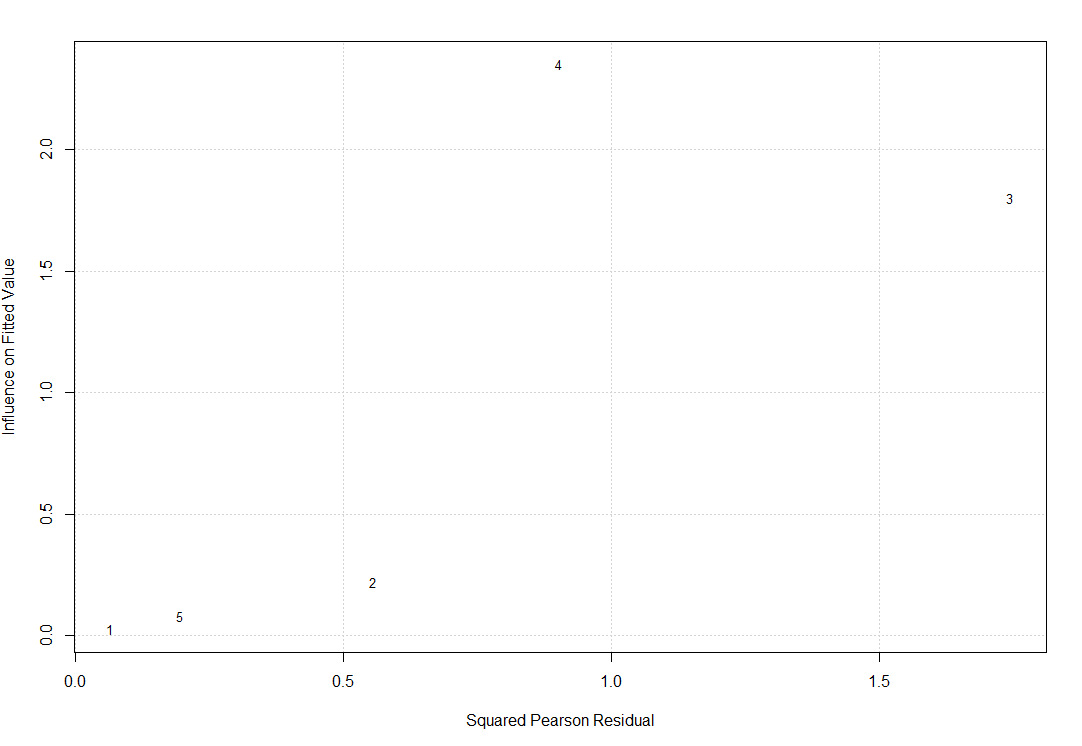


**Fig. A19. Contour enhanced funnel plot for PFS at 3 months (pre-treated patients).**

**Diagnostics for influential studies**

Examination of the externally standardized residuals showed that studies in line 4 ‘Hensley 2009: Sunitinib’ and line 3 ‘’Duska 2014: Ixabepilone’ have the highest z-value of 2 and -2, respectively. Both studies are potentially influential for the meta-analysis.

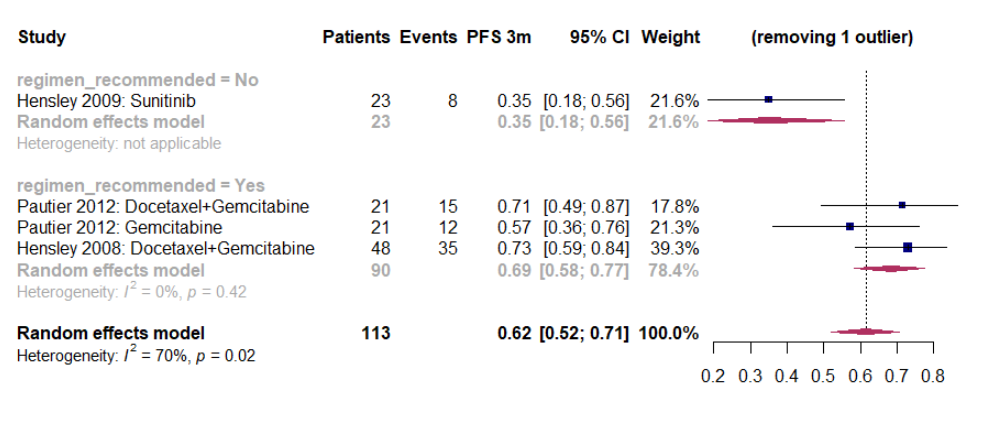
The Baujat plot in fig. A20 projects line with number 3 on the upper right corner. ‘Duska 2014: Ixabepilone’ seems to contribute to the highest amount of heterogeneity (based on squared Pearson residuals) and to be the second most influential (fitted values).



**Fig. A20. Baujat plot of the treatment arms for PFS at 3 months (pre-treated patients).**

**Sensitivity meta-analysis**

A sensitivity meta-analysis was performed to focus on the impact of ‘Duska 2014: Ixabepilone’ for the random-effects model.



**Fig. A21. Forest plot for PFS at 3 months excluding ‘Duska 2014: Ixabepilone’ (pre-treated patients).**

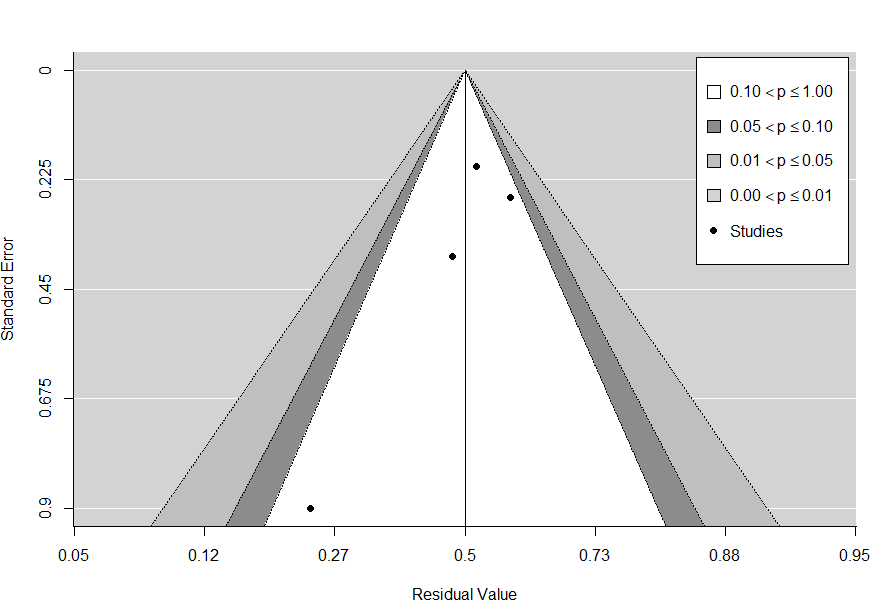
There was a big increase in the overall effect size from 0.53 (0.39 – 0.66) to 0.62 (0.52 – 0.71). The overall heterogeneity was reduced to 70% p = 0.02 (from 83% p < 0.01). Study was removed from the non-recommended group of drugs. The recalculated estimate for this group with one remaining study was 0.35 (0.18 – 0.56) from 0.23 (0.10 – 0.44). Based on the omnibus chi-square test for moderators, there was still statistically significant difference between the subgroups (p < 0.01). Findings indicated that the excluded study was influential for the overall PFS estimate. Results here were not robust, which was expected because of the only 5 available drug regimens (long database).

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

Again, the EORTC database was limited to 4 trials covering 5 different therapeutic combinations.

**Risk of bias assessment**

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



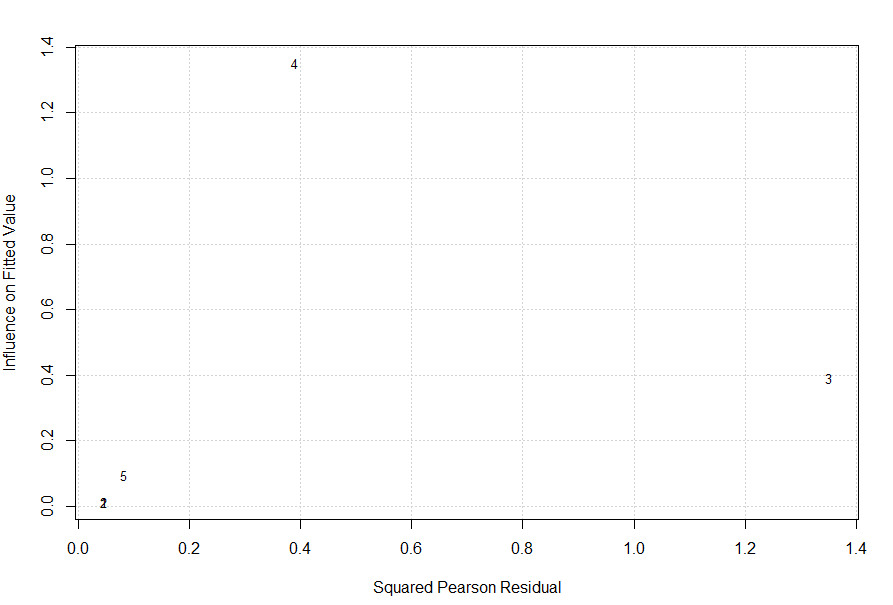
**Fig. A22. Contour enhanced funnel plot for PFS at 6 months (pre-treated patients).**

We also performed the 2 formal tests. The rank correlation test produced a p-value of 0.02 and the Egger’s regression test a p-value of 0.20. There is high risk of publication bias (funnel plot asymmetry) based on the rank correlation test.

**Diagnostics for influential studies**

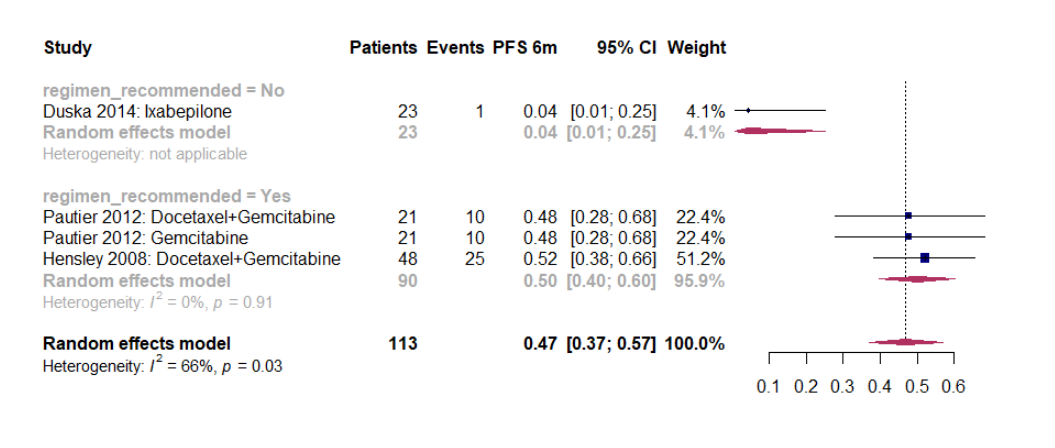
The externally standardized residuals of the studies were calculated. Line 3 and 4 of studies ‘Duska 2014: Ixabepilone’ and ‘Hensley 2009: Sunitinib’ had the highest z-values of 1.32 and -1.32, respectively. The second treatment arm: ‘Hensley 2009: Sunitinib’ was the most influential according to a series of influential diagnostics.

The baujat plot (fig. A23) projects this study to the middle right side. This study (number 3) is the 2nd most influential based on the fitted values on the y-axis and contributes the most to heterogeneity (squared Pearson residuals).



**Fig. A23. Baujat plot of the treatment arms for PFS at 6 months (pre-treated patients).**

**Sensitivity meta-analysis**



**Fig. A24. Forest plot for PFS at 6 months excluding ‘Hensley 2009: Sunitinib’ (pre-treated patients).**

We performed a sensitivity meta-analysis removing ‘Hensley 2009: Sunitinib’ from the database. We fit a DerSimonian and Laird random-effects model (forest plot in fig. A24). The overall effect size increased to 0.47 (0.37 – 0.57) from 0.42 (0.34 – 0.52). Overall heterogeneity was 66% (p = 0.03). Study was removed from the non-recommended group of drugs. The estimated effect size of this subgroup decreased to 0.04 (0.01 – 0.25) from 0.13 (0.05 – 0.28). Based on the omnibus chi-square test for moderators, the difference between the subgroups remained highly statistically significant (p < 0.01). Findings were more robust compared to pretreated population at 3 months but still not that much robust because of the small number of studies.

# References

1 Liberati A, Altman DG, Tetzlaff J, *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**: e1–34.

2 Tap WD, Papai Z, Van Tine BA, *et al.* Doxorubicin plus evofosfamide versus doxorubicin alone in locally advanced, unresectable or metastatic soft-tissue sarcoma (TH CR-406/SARC021): an international, multicentre, open-label, randomised phase 3 trial. *Lancet Oncol* 2017; **18**: 1089–103.

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

4 DerSimonian R, Nan Laird. Meta-Analysis in Clinical Trials Revisited: HHS Public Access. *Contemp Clin Trials* 2015; **45**: 139–45.

5 Borenstein M, Hedges L V., Higgins JPT, Rothstein HR. Introduction to Meta-Analysis. John Wiley & Sons, 2011 https://books.google.be/books/about/Introduction\_to\_Meta\_Analysis.html?id=JQg9jdrq26wC&source=kp\_cover&redir\_esc=y (accessed July 28, 2019).

6 Casali PG, Abecassis N, Aro HT, *et al.* Soft tissue and visceral sarcomas: ESMO-EURACAN Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol* 2018; **29**: iv51–67.

7 Viechtbauer W, Cheung MW-L. Outlier and influence diagnostics for meta-analysis. *Res Synth Methods* 2010; **1**: 112–25.

8 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.

9 Sterne JAC, Egger M. Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *J Clin Epidemiol* 2001; **54**: 1046–55.

10 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.

11 Begg CB, Mazumdar M. Operating Characteristics of a Rank Correlation Test for Publication Bias. *Biometrics* 1994; **50**: 1088–101.

12 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.

13 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.

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

15 Miller JJ. The inverse of the freeman-tukey double arcsine transformation. *Am Stat* 1978; **32**: 138–138.