Published in final edited form as:
Cancer Res. 2016 August 15; 76(16): 4627–4636. doi:10.1158/0008-5472.CAN-15-3455.

Design and Reporting of Targeted Anticancer Preclinical Studies: A Meta-Analysis of Animal Studies Investigating Sorafenib Antitumor Efficacy

James Mattina[1], **Nathalie MacKinnon**[1], **Valerie C. Henderson**[1], **Dean Fergusson**[2], and **Jonathan Kimmelman**[1]

[1]Studies of Translation, Ethics and Medicine (STREAM), Biomedical Ethics Unit, McGill University, 3647 Peel Street H3A 1X1, Montréal, Québec, Canada

[2]Department of Clinical Epidemiology, Ottawa Hospital Research Institute, 725 Parkdale Avenue K1H 8L6, Ottawa, Ontario, Canada

Abstract

The validity of preclinical studies of candidate therapeutic agents has been questioned given their limited ability to predict their fate in clinical development, including due to design flaws and reporting bias. In this study, we examined this issue in depth by conducting a meta-analysis of animal studies investigating the efficacy of the clinically approved kinase inhibitor, sorafenib. MEDLINE, Embase, and BIOSIS databases were searched for all animal experiments testing tumor volume response to sorafenib monotherapy in any cancer published until April 20, 2012. We estimated effect sizes from experiments assessing changes in tumor volume and conducted subgroup analyses based on prespecified experimental design elements associated with internal, construct, and external validity. The meta-analysis included 97 experiments involving 1761 animals. We excluded 94 experiments due to inadequate reporting of data. Design elements aimed at reducing internal validity threats were implemented only sporadically, with 66% reporting animal attrition and none reporting blinded outcome assessment or concealed allocation. Anticancer activity against various malignancies was typically tested in only a small number of model systems. Effect sizes were significantly smaller when sorafenib was tested against either a different active agent or combination arm. Trim and fill suggested a 37% overestimation of effect sizes across all malignancies due to publication bias. We detected a moderate dose-response in one clinically approved indication, hepatocellular carcinoma, but not in another approved malignancy, renal cell carcinoma, or when data were pooled across all malignancies tested. In support of other reports, we found that few preclinical cancer studies addressed important internal, construct and external validity threats, limiting their clinical generalizability. Our findings reinforce the need to improve guidelines for the design and reporting of preclinical cancer studies.

Corresponding Author: Jonathan Kimmelman, Studies of Translation, Ethics and Medicine (STREAM), Biomedical Ethics Unit, McGill University, 3647 Peel Street H3A 1X1, Montréal, Québec, Canada. Phone: 514-398-6980; Fax: 514-398-8349; jonathan.kimmelman@mcgill.ca.

Conflict of interest statement: The authors disclose no potential conflicts of interest.

Introduction

Several recent reports have raised questions about the reproducibility of preclinical studies in general—as well as in particular—in cancer (1, 2). Various commentators have posited different reasons for this problem. One is the use of small sample sizes, which lead to high random variation of results (3). Another is the use of methods, like non-blinding of outcome assessment, that introduce validity threats (4). Such practices, when coupled to publication bias, would lead to especially exaggerated and non-reproducible estimates of effect sizes.

In a previous report, we investigated design, reporting and outcomes for tumor volume experiments contained within preclinical studies of the anticancer drug, sunitinib (5). We found that design practices that reduce the threat of bias and random variation, such as outcome assessment blinding, were rarely implemented. Our analysis suggested that effect sizes were inflated when sunitinib was tested in only one model system, but not necessarily when researchers failed to implement measures like randomization. We also reported evidence that effect sizes may have been overestimated by 45% due to publication bias. Last, we found little relationship between sunitinib properties in preclinical studies, and those that have been observed in humans. For instance, we were unable to detect a dose-response effect when we pooled all studies, and all malignancies responded significantly to sunitinib in preclinical studies, even though not all malignancies have responded in clinical trials.

However, this previous study concentrated on a single drug, and was not based on a prespecified protocol. The extent to which our findings generalize to other preclinical cancer studies is unclear. To explore the generalizability of our findings, we undertook a nearly identical systematic review of all preclinical monotherapy studies for the drug, sorafenib. Sorafenib (Nexavar®, BAY 43-9006) is, like sunitinib, a multikinase inhibitor. It is approved for use in renal cell carcinoma (RCC) (6), hepatocellular carcinoma (HCC) (7), and thyroid cancer (8). This drug was chosen because it has been tested against a large number of different malignancies—many of which have been tested in trials as well. It also provides years worth of follow-up preclinical testing. In this report, we survey experimental design parameters for sorafenib preclinical studies and examine whether design features correlated with estimated effect sizes.

Materials and Methods

Literature Search

Studies were identified by searching MEDLINE, Embase, and BIOSIS databases on April 20, 2012 for trials using these search terms: “sorafenib,” or “Nexavar,” or variations on “BAY 43-9006,” and MeSH terms including “preclinical,” “animals,” or search terms for commonly used animal models. The full search strategy, adapted from Hooijmans *et al.* (9) and de Vries *et al.* (10) can be found in Supplementary Text 1. A PRISMA flow diagram (11) can be found in Supplementary Figure 1.

Study Selection

Inclusion criteria at the study level were a) primary data, b) full-text articles b) English language, c) investigated anticancer efficacy, d) measured a treatment effect in live, nonhuman animals, e) administered sorafenib monotherapy as a comparator or treatment arm. For inclusion at the experiment level and quantitative meta-analysis, additional criteria were f) tested sorafenib against a control arm (e.g. vehicle), g) measured variance as standard deviation of the mean (SD) or standard error of the mean (SEM), h) evaluated drug effect on primary tumor volume, and i) measured baseline tumor volume plus at least one common time measurement between control and treatment arms.

Data Extraction

We extracted experimental design elements derived from a prior systematic review of preclinical research guidelines (12). These included the following at the study level: the names of the authors, the month and year of publication, the country associated with the corresponding author, the funding source(s), the conflict of interest statement, the molecular or physiological rationale for the experiment, and the authors' recommendation of sorafenib in the clinical setting as monotherapy or in combination therapy.

The design elements extracted at the experiment level included the following: sample size of each arm, randomization of treatment allocation, blinding of outcome assessment, inferential statistical test used, removal of animals throughout experiment, species, sex, weight, age, strain, immune status, disease modeled, disease stage, method of tumor initiation, transplantation site, transplantation size, transplant identity, drug administration schedule, administration method, days from disease induction to treatment, day discontinuation of treatment, and presence of combination and comparator arms. We captured the treatment effect at baseline, day 14 (or closest point), last time point, last common time point between control and treatment arms, as well as the standard deviation of the mean (SD) or standard error of the mean (SEM).

We extracted experiments that measured treatment effect as tumor volume (usually in units of mm^3) or used a reasonable proxy for tumor volume, including the following: caliper measurement (mm^3), tumor weight (mg), optical measurement ($photons \cdot s^{-1}$), and fold change in tumor volume between control and treatment arms. To account for the heterogeneity of scales in tumor volume measurements, we calculated effect sizes as standardized mean differences (SMDs) using Hedges' *g*. The Hedges' *g* statistic measures effect sizes in terms of the variability observed within individual studies, producing a standardized measure of treatment effect and allowing for the combination of results (13). We extracted, but did not analyze survival information due to the paucity of data reported (Table 1).

Graphical data were extracted using GraphClick digitizer software (Arizona Software). All extractions were performed by NM. After piloting, we identified eight extraction items that were prone to high inter-rater variability (Supplementary Text 2). These items were double-coded independently by JM and reconciled by discussion.

Meta-analysis

We calculated the effect sizes as SMDs using Hedges' *g* and 95% confidence intervals. We used the statistical software, OpenMeta[Analyst] (14) to calculate pooled effect sizes using the DerSimonian and Laird random-effects model (15) and to assess heterogeneity of data via I^2 statistics (16). Statistical significance was set at a *p*-value of 0.05; because of multiplicities, all testing was exploratory. We did not prospectively register a protocol for this meta-analysis; however, except where noted, hypothesis testing was prespecified and we followed the methods used in our previous meta-analysis of the anticancer drug, sunitinib (5).

For experiments testing multiple doses of sorafenib, we averaged the outcomes and created a pooled effect size for each experiment (except in dose-response analyses). Publication bias was evaluated using funnel plots (17) with Duval and Tweedie's trim and fill method of estimating missing studies and adjusting the point estimate (18) using Comprehensive Meta Analyst software (19). Funnel plots take advantage of the fact that smaller studies are prone to large random variation. Under-representation of smaller studies showing non-positive effects can suggest publication bias.

For the dose-response curves, we only evaluated experiments using continuous dosing schedules and measuring tumor volume at a fixed time point of 14 days after onset of dosing (±3 days allowed, Supplementary Figure 2B). We excluded all other experiments from this analysis because dosing schedule and time point choice would be expected to correlate with effect sizes.

Results

Study Characteristics

Our search captured 105 studies containing 191 experiments assessing tumor volume response to sorafenib monotherapy. Although all studies were included in qualitative analyses (Table 1), only 65 studies containing 97 experiments were included in our meta-analysis. A total of 94 experiments (49%) were excluded because they did not report elements required for our quantitative tests (e.g. sample size, a measure of dispersion or baseline tumor volume), 44 of which were reported in a single study (20). The 97 included experiments used 1761 animals, 96% of which were mice. The mean duration of experiments used in quantitative meta-analysis was 21 days (range 3–55 days). Anticancer efficacy experiments relied heavily on human xenograft models of disease (95%). Average sample size in each experiment was 7.74 and 7.78 in treatment and control arms, respectively (range 3–20 for both). There was high heterogeneity of data across all studies (I^2=79%) (21). Most studies (98%) were published after sorafenib had received regulatory approval, reflecting the continued exploration of activity against various malignancies and delays in the publication process.

Experimental Practices

Several bodies have called for implementation of a suite of practices in preclinical testing, including randomization and blinding (22–26). A systematic review of preclinical design guidelines identified a consensus set of practices for improving clinical generalizability (12). We examined the reported implementation of these practices in sorafenib studies.

Experimental practices aimed at minimizing bias and strengthening causal inferences (internal validity) varied. Concealed allocation and blinded outcome assessment were never reported as used. Of the experiments in our sample, 10% evaluated dose-response (≥3 doses) of sorafenib. Moreover, 66% of experiments addressed, exhaustively or briefly, the attrition of animals during experiments.

Design elements aimed at maximizing the correspondence between experimental setup and clinical scenarios (construct validity) were also variable. Key parameters identified in preclinical study design guidelines include matching age of animals to patients, matching sex, matching stage of disease, and confirming mechanism of action. Studies relied disproportionately on younger, female animals with less advanced disease (Table 2)—variables that probably do not match most clinical scenarios. However, most studies (79%) probed for molecular or physiological evidence of mechanism of action.

Many guidelines recommend replication in different models of disease to rule out the possibility that treatment effects are attributable to idiosyncrasies in model systems (external validity) (12). We used an index of external validity first by counting the number of species and models used per malignancy. Hepatocellular carcinoma and high-grade glioma experiments employed the greatest variety of species (n=2) and models (n=2), yet as described below, these malignancies did not show significantly smaller effect sizes (Fig. 1A). Next (and on an *ad hoc* basis), we created a new index of external validity by pooling all graft studies for a malignancy type and determining the number of different cell lines used to test activity. We examined whether malignancies that were tested in more model systems tended to show more modest effect sizes. Most malignancies tested sorafenib against one or two tumor cell lines. Although malignancies that tested sorafenib using a single representative cell line (n=6) seemed to show larger effect sizes than those testing in more than one model, this was not significant (Fig. 1B).

Effect Sizes in Preclinical Studies

Effect sizes in experiments, pooled by indication are reflected in Figure 2. The mean effect size across all malignancies was −2.396 (95% CI, −2.682, −2.110). From the 97 included experiments, 76.3% reached statistical significance (p<0.05, Supplementary Figure 3) and 61% of papers concluded by recommending clinical testing. Each pooled malignancy demonstrated significant anticancer activity, except pancreatic cancer (n=2) and squamous cell carcinoma (n=2). Though a quantitative analysis was not possible at this time, malignancies that are known to respond clinically (e.g. RCC (6, 27)) did not suggest substantially larger preclinical pooled effect sizes (Fig. 2) than malignancies that show minimal clinical response to sorafenib monotherapy (e.g. melanoma (28), non-small-cell lung carcinoma (29–31), ovarian (32, 33), and breast cancer (34, 35)). Thyroid carcinoma is also approved, but tumor volume experiments in this indication were missing measurements of baseline tumor volume or variance and were excluded from quantitative meta-analysis.

We performed an exploratory analysis examining whether any experimental design parameters described above corresponded with smaller, and thus likely, more realistic effect sizes. With respect to internal validity practices, there were no clear trends between design and effect size (Fig. 3A). For construct validity, experiments that tested sorafenib as monotherapy against an active comparator, or against a sorafenib-containing combination showed significantly smaller effect sizes than experiments testing against only an inactive control arm (Fig. 3B). Furthermore, experiments that reported a conflict of interest showed significantly smaller effect sizes than those declaring no conflict of interest (Fig. 3B).

Evidence of Publication Bias

One possible explanation for the preponderance of strongly positive studies was publication bias. We constructed funnel plots and performed trim and fill analysis to explore this possibility in our pooled sample. The asymmetric funnel plot for all malignancies (Fig. 4A) suggests the presence of publication bias. Trim and fill analysis suggests a 37% overestimation of effect size across all malignancies, with an adjusted SMD estimate of −1.753 (95% CI −2.073, −1.433) compared to an unadjusted SMD of −2.396 (95% CI −2.682, −2.110). We performed similar analyses for HCC (Fig. 4B) and RCC (Fig. 4C)—the two malignancies for which we had the greatest volume of experiments (n=29 and n=17, respectively). HCC showed no significant suggestion of publication bias. The analyses suggested a 25% overestimation of effect size for RCC, although this was not significant and limited by sample size.

Dose-response Effects

In our sample, ten percent of the experiments performed dose-response curves (≥3 doses) for sorafenib. There is some evidence suggesting dose-response effects in human beings, although these studies are not decisive (36–39). However, preclinical studies that tested dose-response internally demonstrated an effect (Figs. 5B and 5C). As a simple measure of the ability of pooled preclinical studies in our sample to demonstrate causal relationships, we tested for whether we could detect dose-response effects if all eligible experiments (n=91), as well as the indications with the largest volume of experiments (HCC, n=28 and RCC, n=17), were pooled. Using a standardized time point of 14 days after sorafenib administration and restricting our dataset to continuous (daily) dosing schedules, we did not observe a dose-response relationship across all malignancies (p=0.09) (Fig. 5A). Considering the subsets of approved malignancies, HCC experiments showed a moderate dose-response (p<0.001, R^2=0.35) (Fig. 5B) while RCC experiments did not (p=0.86) (Fig. 5C).

Discussion

Preclinical efficacy experiments are typically cited to justify the initiation of clinical trials. However, choice of models, experimental setup, and reporting practices may limit their clinical generalizability. Our report builds on previous findings that experimental practices in preclinical cancer research do not adequately attend to the effects of random variation, bias, and non-publication.

As in our previous study (5), we found that many experiments are reported so poorly that they are almost impossible to interpret. For instance, more than a third of our original sample could not be included in the meta-analysis due to missing information on sample size, measure of dispersion or baseline tumor volume. Similarly, we found limited attention to internal validity threats, as indicated by the general non-implementation and reporting of design elements such as concealed allocation, blinded outcome assessment, and animal attrition. With respect to construct validity, researchers generally relied on young, immunocompromised and female mice, as they had for sunitinib (5). In this study and in the previous one, however, we did not detect exaggerated effect sizes in studies harbouring internal or construct validity threats, as others have (40, 41). Our analysis suggested that experimental effect sizes were significantly smaller when sorafenib was tested against active comparators and/or combination arms—a finding that would be consistent with bias (since the purpose of such studies is to demonstrate that another drug or combination is even more effective) but that contradicts our sunitinib results (5).

Our analysis is suggestive of biases in reporting of sorafenib preclinical studies. First, 76.3% of studies were statistically significant—a proportion that is surprising, given that the mean sample size per arm was small (n=7.76). As with sunitinib, almost all malignancies demonstrated statistically significant activity—the two that did not trended strongly towards positivity. If all malignancies respond to sorafenib, the value for trial planning of the type of *in vivo* testing used in experiments analyzed here is doubtful. Third, our trim and fill analysis suggested an overestimation of effect size due to publication bias across malignancies that is similar to what we observed for sunitinib (5). Our analysis did not find a strong dose-response relationship for pooled malignancies—nor for one of the malignancies currently approved for monotherapy. Fourth, similar to the results our preclinical sunitinib report (5), our external validity analysis suggested that testing in more model systems—the number of grafts, species, and models used—results in more realistic (i.e. smaller) pooled effect sizes within malignancies, although this trend was non-significant. Last, our findings do not suggest a clear relationship between preclinical effect sizes and clinical outcomes across malignancies, though a more formal analysis including clinical effect sizes is still needed.

Our analysis and inferences about effect sizes have many limitations, not least of which is the hazard of combining effect sizes from an extremely heterogeneous sample of experiments. For example, toxicity of drug at high doses may have dampened the ability of xenograft studies to detect dose-responses (although this fails to explain why they are consistently reported internally) and cell line heterogeneity may mask dose-effects within malignancies. Although the administered dose was always reported, the lack of reported drug exposure data threatens the construct validity of the experiments in our sample. Second, our analysis was focused on only *in vivo* experiments embedded within preclinical reports. It is possible that tumor volume curves should only be interpreted in the context of additional mechanistic, pharmacokinetic, or *in vitro* experiments within reports. Third, our systematic review relied on what was published and reported in studies. It is possible studies may have used methodologies, like randomization, and not reported them. Fourth, our systematic review concerns a single drug. Although our findings are consistent with observations reported elsewhere (4, 5, 42, 43), it is possible more robust dose-response curves, or a better relationship between clinical and preclinical effects would be apparent with other drugs. Fifth, our analysis only captures studies published before April 2012, however, we believe that extending our results to the current date would not reveal vastly different treatment outcomes or quality of reporting. Last, our results may reflect problems with using human xenograft tumor growth curves to make clinical inferences—particularly for a drug like sorafenib, which shows cytostatic properties in clinical trials (44, 45). Data suggest that tumor shrinkage may not be a suitable efficacy endpoint for sorafenib; time-to-event data, including prolongation of progression-free survival (PFS) and overall survival (OS), indicate benefits from tumor stabilization despite modest radiographic response in pivotal trials (6, 7). However, our analysis does not include survival data due to the scarcity of reported survival curves in our sample (Table 1). We also note that mean effect sizes observed in sorafenib preclinical studies were much greater than those observed for sunitinib (−2.396 [95% CI, −2.682, −2.110] vs. −1.826 [95% CI, −2.052, −1.601], respectively), a drug associated with high objective response rates in trials. While the usefulness and reproducibility of the xenograft model have been questioned (46–48), many support its use in preclinical studies (49–51).

Our findings contribute to the literature on preclinical design and reporting in cancer, and reinforce our exploratory analysis for sunitinib (5). They also suggest that researchers—and physicians prescribing approved drugs off-label—should be cautious about using tumor curves to infer clinical value. Many xenograft studies do not adhere to basic tenets of reporting, such as describing sample sizes; few implement widely discussed design elements like blinding. It might be objected that cancer represents a “hard endpoint”—and hence is less susceptible to bias than other disease realms. However, measurements of tumor volume, assessments of moribundity for survival curves, or choices of whether to include anomalous measurements involve judgment, and just as in clinical research, such judgments can be affected by bias. We encourage the cancer research community to pursue a sustained discussion of guidelines for experimental setup and results reporting in preclinical research. We also encourage referees to scrutinize manuscripts for reporting. Above all, our findings suggest possibilities for reducing some of the burden and cost associated with unsuccessful translation efforts.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding statement: This work was funded by CIHR (EOG111391).

We thank Amanda Hakala and Benjamin Carlisle for their assistance with this manuscript.

References

1. Begley CG, Ellis LM. Drug development: Raise standards for preclinical cancer research. Nature. 2012; 483:531–3. [PubMed: 22460880]
2. Morrison SJ. Time to do something about reproducibility. eLife. 2014:3.
3. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, et al. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013; 14:365–76. [PubMed: 23571845]
4. Amarasingh S, Macleod MR, Whittle IR. What is the translational efficacy of chemotherapeutic drug research in neuro-oncology? A systematic review and meta-analysis of the efficacy of BCNU and CCNU in animal models of glioma. J Neurooncol. 2009; 91:117–25. [PubMed: 18813876]

5. Henderson VC, Demko N, Hakala A, MacKinnon N, Federico CA, Fergusson D, et al. A meta-analysis of threats to valid clinical inference in preclinical research of sunitinib. eLife. 2015; 4:e08351. [PubMed: 26460544]

6. Escudier B, Eisen T, Stadler WM, Szczylik C, Oudard S, Siebels M, et al. Sorafenib in advanced clear-cell renal-cell carcinoma. N Engl J Med. 2007; 356:125–34. [PubMed: 17215530]

7. Llovet JM, Ricci S, Mazzaferro V, Hilgard P, Gane E, Blanc JF, et al. Sorafenib in advanced hepatocellular carcinoma. N Engl J Med. 2008; 359:378–90. [PubMed: 18650514]

8. Brose MS, Nutting CM, Jarzab B, Elisei R, Siena S, Bastholt L, et al. Sorafenib in radioactive iodine-refractory, locally advanced or metastatic differentiated thyroid cancer: a randomised, double-blind, phase 3 trial. Lancet. 2014; 384:319–28. [PubMed: 24768112]

9. Hooijmans CR, Tillema A, Leenaars M, Ritskes-Hoitinga M. Enhancing search efficiency by means of a search filter for finding all studies on animal experimentation in PubMed. Lab Anim. 2010; 44:170–5. [PubMed: 20551243]

10. Vries RB, Hooijmans CR, Tillema A, Leenaars M, Ritskes-Hoitinga M. A search filter for increasing the retrieval of animal studies in Embase. Lab Anim. 2011; 45:268–70. [PubMed: 21890653]

11. Moher D, Liberati A, Tetzlaff J, Altman DG. Group Prisma. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med. 2009; 6:e1000097. [PubMed: 19621072]

12. Henderson VC, Kimmelman J, Fergusson D, Grimshaw JM, Hackam DG. Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. PLoS Med. 2013; 10:e1001489. [PubMed: 23935460]

13. Durlak JA. How to select, calculate, and interpret effect sizes. J Pediatr Psychol. 2009; 34:917–28. [PubMed: 19223279]

14. Wallace BC, Schmid CH, Lau J, Trikalinos TA. Meta-Analyst: software for meta-analysis of binary, continuous and diagnostic data. BMC Med Res Methodol. 2009; 9:80. [PubMed: 19961608]

15. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials. 1986; 7:177–88. [PubMed: 3802833]

16. Higgins JP, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002; 21:1539–58. [PubMed: 12111919]

17. Light, RJ., Pillemer, DB. Summing up: the science of reviewing research. Cambridge, Mass: Harvard University Press; 1984.

18. Duval S, Tweedie R. Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. Biometrics. 2000; 56(2):455–63. [PubMed: 10877304]

19. Dietz G, Dahabreh IJ, Gurevitch J, Lajeunesse MJ, Schmid CH, Trikalinos TA, et al. OpenMEE: Software for Ecological and Evolutionary Meta-Analysis. Computer program. 2015

20. Keir ST, Maris JM, Lock R, Kolb EA, Gorlick R, Carol H, et al. Initial testing (stage 1) of the multi-targeted kinase inhibitor sorafenib by the pediatric preclinical testing program. Pediatric blood and cancer. 2010; 55:1126–1133. [PubMed: 20672370]

21. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003; 327:557–60. [PubMed: 12958120]

22. Begley CG. Six red flags for suspect work. Nature. 2013; 497:433–4. [PubMed: 23698428]

23. Fisher M, Feuerstein G, Howells DW, Hurn PD, Kent TA, Savitz SI, et al. Update of the stroke therapy academic industry roundtable preclinical recommendations. Stroke. 2009; 40:2244–50. [PubMed: 19246690]

24. Landis SC, Amara SG, Asadullah K, Austin CP, Blumenstein R, Bradley EW, et al. A call for transparent reporting to optimize the predictive value of preclinical research. Nature. 2012; 490:187–91. [PubMed: 23060188]

25. Stroke Therapy Academic Industry Roundtable. Recommendations for standards regarding preclinical neuroprotective and restorative drug development. Stroke. 1999; 30:2752–8. [PubMed: 10583007]

26. Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. PLoS Biol. 2010; 8:e1000412. [PubMed: 20613859]

27. Jonasch E, Corn P, Pagliaro LC, Warneke CL, Johnson MM, Tamboli P, et al. Upfront, randomized, phase 2 trial of sorafenib versus sorafenib and low-dose interferon alfa in patients with advanced renal cell carcinoma. Cancer. 2010; 116:57–65. [PubMed: 19862815]

28. Ott Patrick A, Hamilton A, Min C, Safarzadeh-Amiri S, Goldberg L, Yoon J, et al. A phase II trial of sorafenib in metastatic melanoma with tissue correlates. PLoS One. 2010; 5:e15588. [PubMed: 21206909]

29. Dy GK, Hillman SL, Rowland KM, Molina JR, Steen PD, Wender DB, et al. A front-line window of opportunity phase 2 study of sorafenib in patients with advanced nonsmall cell lung cancer. Cancer. 2010; 116:5686–5693. [PubMed: 21218460]

30. Blumenschein GR, Gatzemeier U, Fossella F, Stewart DJ, Cupit L, Cihon F, et al. Phase II, multicenter, uncontrolled trial of single-agent sorafenib in patients with relapsed or refractory, advanced non-small-cell lung cancer. Journal of Clinical Oncology. 2009; 27:4274–4280. [PubMed: 19652055]

31. Kim ES, Herbst RS, Wistuba II, Lee JJ, Blumenschein GR, Tsao A, et al. The BATTLE trial: personalizing therapy for lung cancer. Cancer discovery. 2011; 1:44–53. [PubMed: 22586319]

32. Matei D, Sill MW, Lankes HA, DeGeest K, Bristow RE, Mutch D, et al. Activity of sorafenib in recurrent ovarian cancer and primary peritoneal carcinomatosis: a gynecologic oncology group trial. Journal of Clinical Oncology. 2011; 29:69–75. [PubMed: 21098323]

33. Bodnar L, Górnas M, Szczylik C. Sorafenib as a third line therapy in patients with epithelial ovarian cancer or primary peritoneal cancer: a phase II study. Gynecologic oncology. 2011; 123:33–36. [PubMed: 21723597]

34. Moreno-Aspitia A, Morton RF, Hillman DW, Lingle WL, Rowland KM, Wiesenfeld M, et al. Phase II trial of sorafenib in patients with metastatic breast cancer previously exposed to anthracyclines or taxanes: North Central Cancer Treatment Group and Mayo Clinic Trial N0336. Journal of Clinical Oncology. 2009; 27:11–15. [PubMed: 19047293]

35. Bianchi G, Loibl S, Zamagni C, Salvagni S, Raab G, Siena S, et al. Phase II multicenter, uncontrolled trial of sorafenib in patients with metastatic breast cancer. Anti-cancer drugs. 2009; 20:616–624. [PubMed: 19739318]

36. Semrad TJ, Gandara DR, Lara. Enhancing the clinical activity of sorafenib through dose escalation: rationale and current experience. Ther Adv Med Oncol. 2011; 3:95–100. [PubMed: 21789159]

37. Strumberg D, Clark JW, Awada A, Moore MJ, Richly H, Hendlisz A, et al. Safety, pharmacokinetics, and preliminary antitumor activity of sorafenib: a review of four phase I trials in patients with advanced refractory solid tumors. Oncologist. 2007; 12(4):426–37. [PubMed: 17470685]

38. Amato RJ, Jac J, Harris P, et al. A phase II trial of intra-patient dose escalated-sorafenib in patients (pts) with metastatic renal cell cancer (MRCC). Journal of Clinical Oncology. 2008; 26(15) abstract.

39. Gore ME, Jones RJ, Ravaud A, et al. Efficacy and safety of intrapatient dose escalation of sorafenib as first-line treatment for metastatic renal cell carcinoma (mRCC). Journal of Clinical Oncology. 2011; 29(15) abstract.

40. Crossley NA, Sena E, Goehler J, Horn J, van der Worp B, Bath PM, et al. Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. Stroke. 2008; 39:929–34. [PubMed: 18239164]

41. Rooke ED, Vesterinen HM, Sena ES, Egan KJ, Macleod MR. Dopamine agonists in animal models of Parkinson's disease: a systematic review and meta-analysis. Parkinsonism Relat Disord. 2011; 17:313–20. [PubMed: 21376651]

42. Hirst TC, Vesterinen HM, Sena ES, Egan KJ, Macleod MR, Whittle IR. Systematic review and meta-analysis of temozolomide in animal models of glioma: was clinical efficacy predicted? Br J Cancer. 2013; 108:64–71. [PubMed: 23321511]

43. Sugar E, Pascoe AJ, Azad N. Reporting of preclinical tumor-graft cancer therapeutic studies. Cancer Biol Ther. 2012; 13:1262–8. [PubMed: 22895077]

44. Abou-Alfa GK, Schwartz L, Ricci S, Amadori D, Santoro A, Figer A, et al. Phase II study of sorafenib in patients with advanced hepatocellular carcinoma. Journal of Clinical Oncology. 2006; 24:4293–4300. [PubMed: 16908937]

45. Ratain MJ, Eisen T, Stadler WM, Flaherty KT, Kaye SB, Rosner GL, et al. Phase II placebo-controlled randomized discontinuation trial of sorafenib in patients with metastatic renal cell carcinoma. Journal of Clinical Oncology. 2006; 24:2505–2512. [PubMed: 16636341]

46. Lorsch JR, Collins FS, Lippincott-Schwartz J. Cell Biology. Fixing problems with cell lines. Science. 2014; 346:1452–3. [PubMed: 25525228]

47. Sharpless NE, Depinho RA. The mighty mouse: genetically engineered mouse models in cancer drug development. Nat Rev Drug Discov. 2006; 5:741–54. [PubMed: 16915232]

48. Olive KP, Tuveson DA. The use of targeted mouse models for preclinical testing of novel cancer therapeutics. Clin Cancer Res. 2006; 12:5277–87. [PubMed: 17000660]

49. Voskoglou-Nomikos T, Pater JL, Seymour L. Clinical predictive value of the in vitro cell line, human xenograft, and mouse allograft preclinical cancer models. Clin Cancer Res. 2003; 9:4227–39. [PubMed: 14519650]

50. Johnson JI, Decker S, Zaharevitz D, Rubinstein LV, Venditti JM, Schepartz S, et al. Relationships between drug activity in NCI preclinical in vitro and in vivo models and early clinical trials. Br J Cancer. 2001; 84:1424–31. [PubMed: 11355958]

51. Kerbel RS. Human tumor xenografts as predictive preclinical models for anticancer drug activity in humans: better than commonly perceived-but they can be improved. Cancer Biol Ther. 2003; 2:S134–9. [PubMed: 14508091]

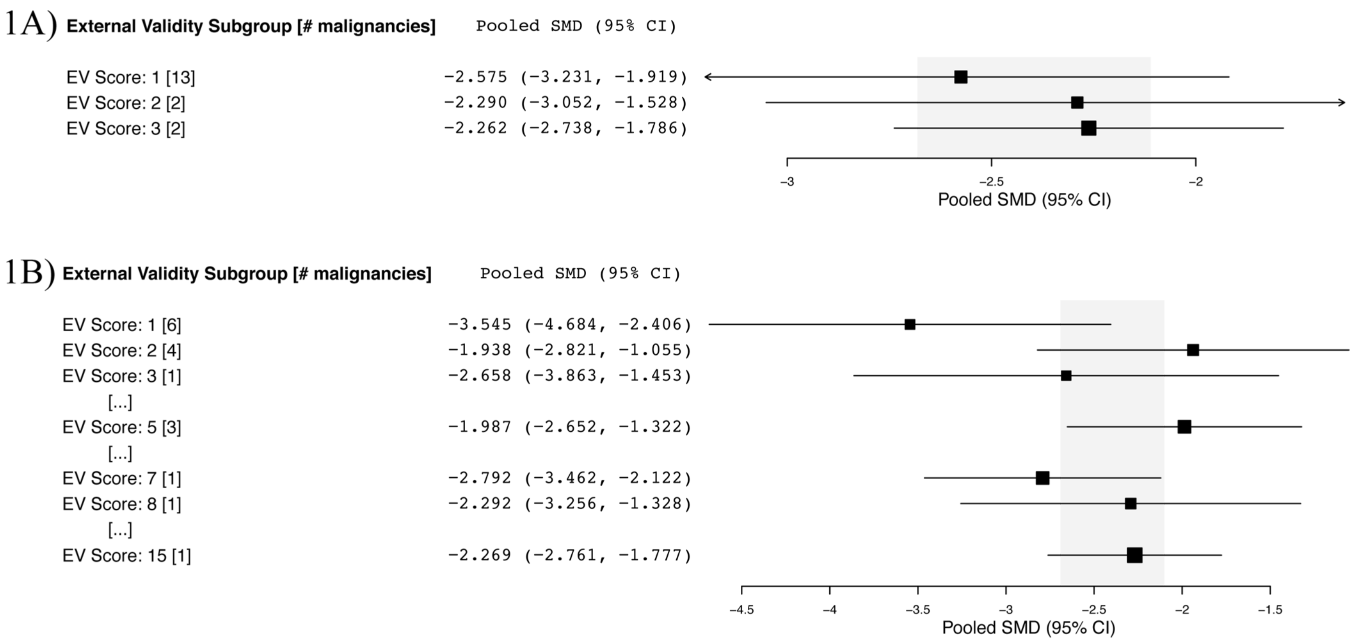

Figure 1. External validity (EV) score per malignancy and subgroup analysis of effect size

A) EV score based on number of species plus disease models used minus one. Hepatocellular carcinoma (*n*=29) and high-grade glioma (*n*=3) experiments employed the greatest variety of species (*n*=2) and models (*n*=2), for an EV score of 3. B) EV score based on number of unique grafts/cell lines used. Hepatocellular carcinoma experiments tested sorafenib against 15 unique grafts.

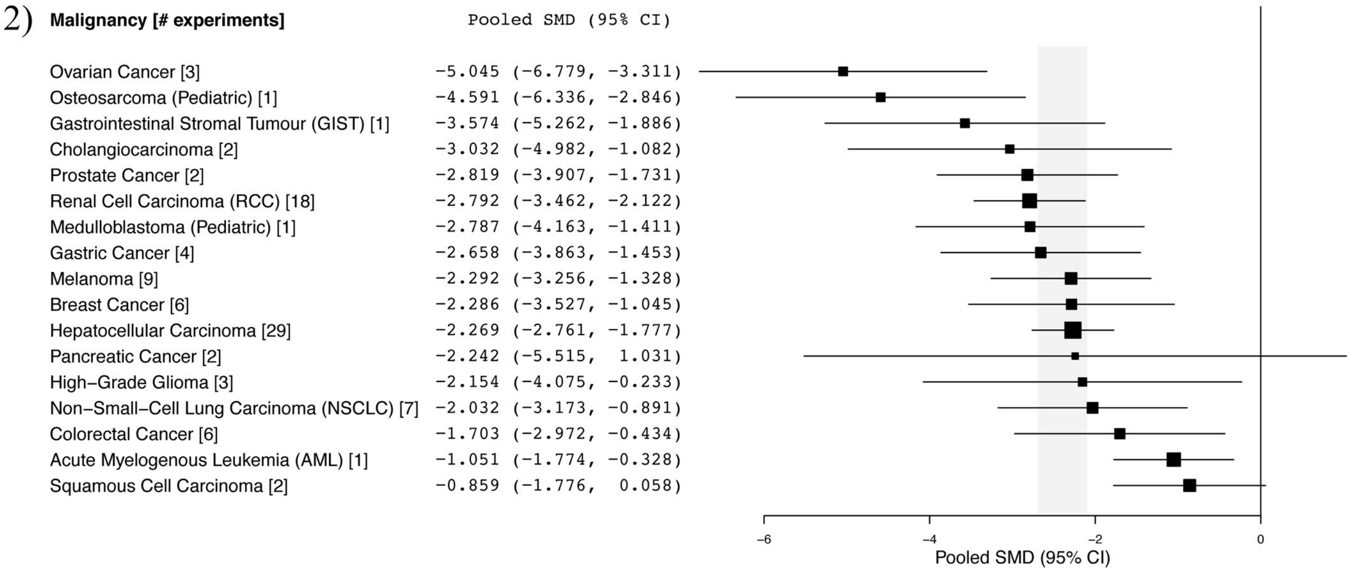

Figure 2. Summary of pooled standardized mean differences (SMDs) per indication

Shaded region indicates overall pooled SMD and 95% CI (−2.396 [−2.682, −2.110]) for all tumor growth experiments (*n*=97).

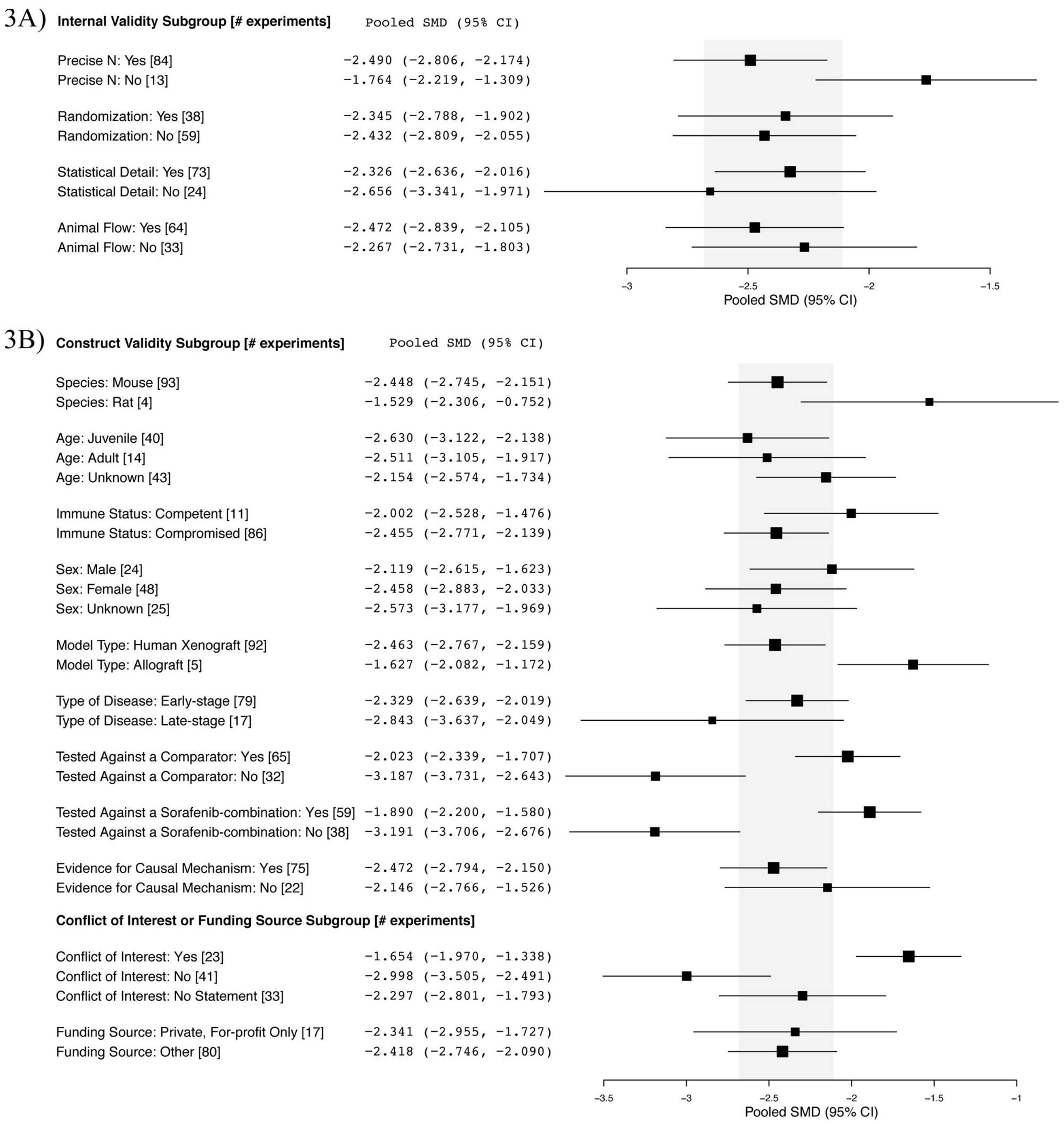

Figure 3. A) Internal validity, B) construct validity, conflict of interest and funding source subgroup analyses

Shaded region denotes the pooled SMD and 95% CI (−2.396 [−2.682, −2.110]) for all tumor growth experiments (*n*=97). The disease stage in one experiment was not reported.

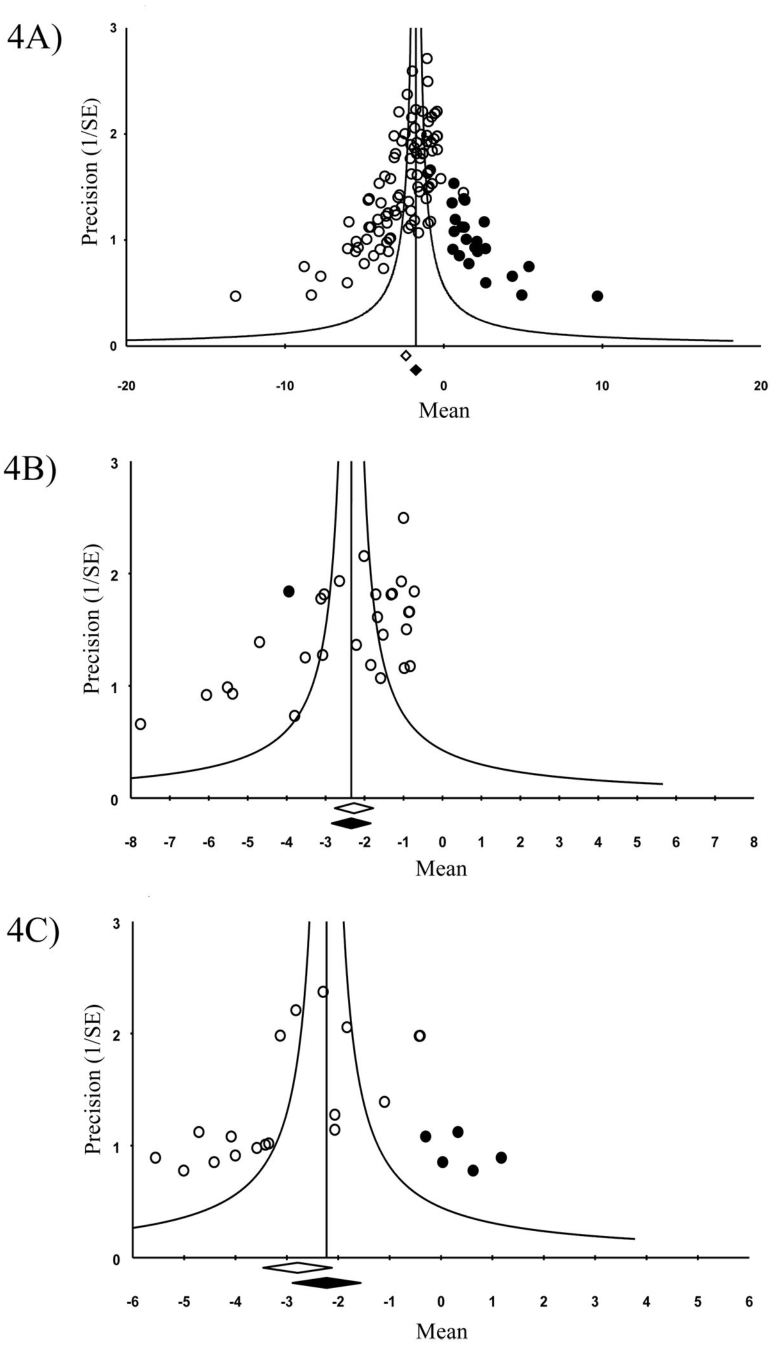

Figure 4. Funnel plots to detect publication bias with trim and fill analysis

A) All eligible tumor growth experiments (*n*=97), B) hepatocellular carcinoma (*n*=29), and C) renal cell carcinoma (*n*=18). Open circles denote data points from included experiments whereas black circles denote “filled” experiments. Open diamond indicates unadjusted SMD whereas black diamond indicates adjusted SMD.

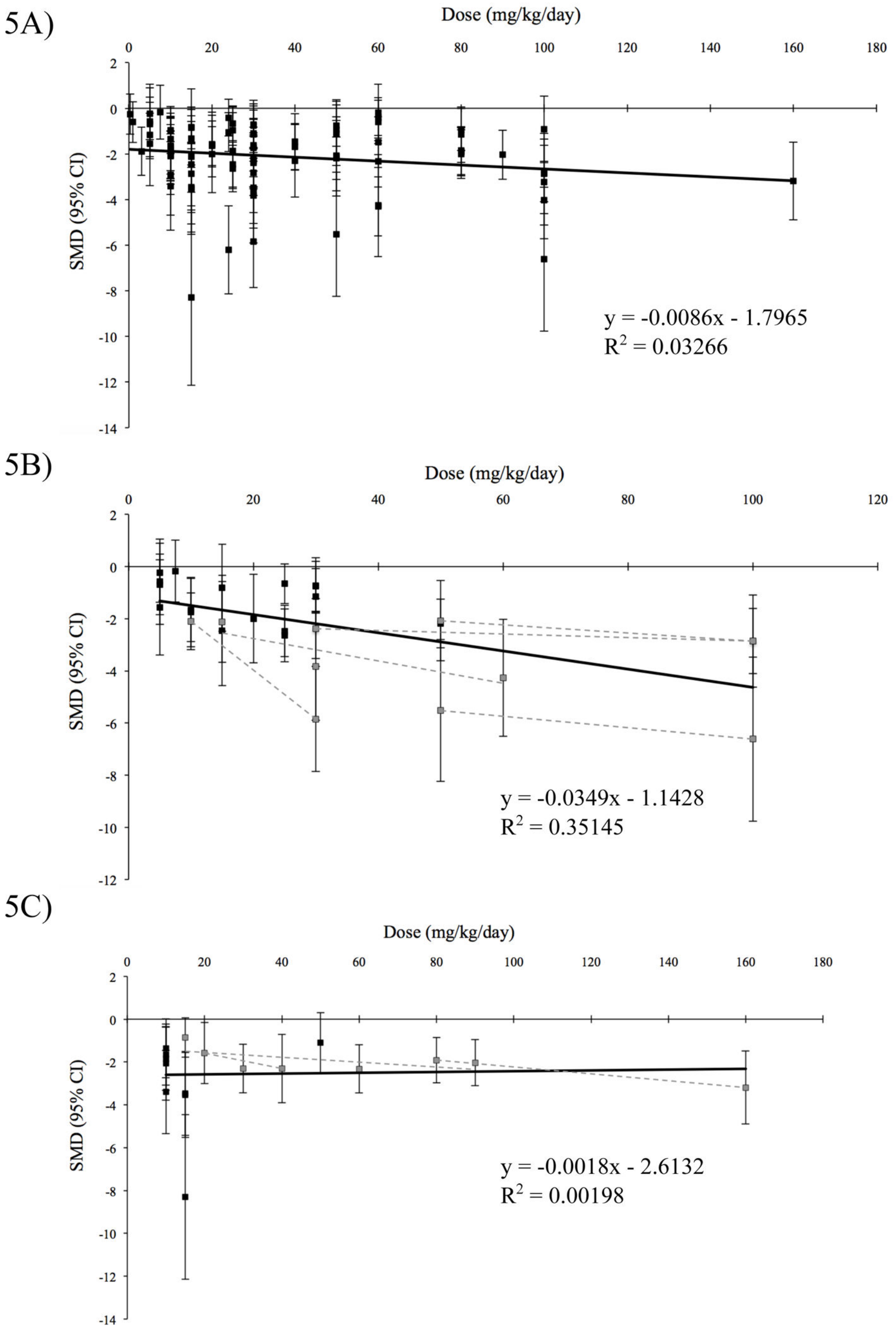

Figure 5. Dose-response curves

A) Experiments from all indications (*n*=88), B) hepatocellular carcinoma (*n*=28), and C) renal cell carcinoma (*n*=17). Only experiments with a common time point of 14 days (±3 days) and continuous dosing were included. Effect sizes were taken from the standardized time point. Dose-response curves within single studies reporting internal dose-response curves (dashed lines) were superimposed for B) hepatocellular carcinoma and C) renal cell carcinoma.

Table 1

Demographics of included studies.

Study-level Demographics		Included Studies (n = 105)
Funding source(s) *	Private, for-profit	26 (25%)
	Private, not for-profit	51 (49%)
	Public	67 (64%)
Recommended clinical testing	Yes	64 (61%)
Reported survival data	Yes	14 (13%)
Publication date	2004–2006	7 (7%)
	2007–2009	28 (27%)
	2010–2012	70 (67%)
Conflict of interest	Conflict of interest declared	23 (22%)
	No conflict of interest	39 (37%)
	No statement	43 (41%)

*Studies may have more than one funding source

Table 2

Descriptive analysis of reported internal and construct validity design elements.

Internal Validity Characteristics		**Included experiments (n = 97)**
Exact sample size given for all groups		84 (87%)
Randomized treatment allocation		38 (39%)
Blinded treatment allocation		0 (0%)
Blinded outcome assessment		0 (0%)
Used specified inferential statistical test		73 (75%)
Addressed animal flow through experiment		64 (66%)
Evaluated dose-response (≥3 doses)		10 (10%)
Construct Validity Characteristics		
Species	Mouse	93 (96%)
	Rat	4 (4%)
Age[*]	Pediatric/Juvenile (≤8wk)	40 (41%)
	Adult (>8wk, <21wk)	14 (14%)
	Aged (≥ 21wk)	0 (0%)
	No data	43 (44%)
Immune status	Immunocompetent	11 (12%)
	Immunocompromised	86 (88%)
Sex	Male	24 (24%)
	Female	48 (49%)
	No data	25 (26%)
Model type	Human xenograft	92 (95%)
	Allograft	5 (5%)
Type of disease[†]	Early-stage (≤200 mm^3) tumor growth	79 (81%)
	Late-stage (>200 mm^3) tumor growth	17 (18%)
Evidence for causal mechanism[‡]	Molecular	37 (38%)
	Physiological	68 (70%)

Coding details can be found in Supplementary Table 1 and Supplementary Table 2.

[*] Age ranges presented for mice. Rats were considered adult at >7 weeks

[†] The disease stage in one experiment was not reported

[‡] Studies may have molecular and physiological evidence for causal mechanism