eLIFE elifesciences.org

RESEARCH ARTICLE

A meta-analysis of threats to valid clinical inference in preclinical research of sunitinib

Valerie C Henderson[1], Nadine Demko[1], Amanda Hakala[1], Nathalie MacKinnon[1], Carole A Federico[1], Dean Fergusson[2], Jonathan Kimmelman[1]*

[1]Studies of Translation, Ethics and Medicine Research Group, Biomedical Ethics Unit, McGill University, Montréal, Canada; [2]Department of Clinical Epidemiology, Ottawa Hospital Research Institute, Ottawa, Canada

Abstract Poor study methodology leads to biased measurement of treatment effects in preclinical research. We used available sunitinib preclinical studies to evaluate relationships between study design and experimental tumor volume effect sizes. We identified published animal efficacy experiments where sunitinib monotherapy was tested for effects on tumor volume. Effect sizes were extracted alongside experimental design elements addressing threats to valid clinical inference. Reported use of practices to address internal validity threats was limited, with no experiments using blinded outcome assessment. Most malignancies were tested in one model only, raising concerns about external validity. We calculate a 45% overestimate of effect size across all malignancies due to potential publication bias. Pooled effect sizes for specific malignancies did not show apparent relationships with effect sizes in clinical trials, and we were unable to detect dose–response relationships. Design and reporting standards represent an opportunity for improving clinical inference.

DOI: 10.7554/eLife.08351.001

***For correspondence:** jonathan.kimmelman@mcgill.ca

Competing interests: The authors declare that no competing interests exist.

Funding: See page 11

Received: 25 April 2015
Accepted: 05 September 2015
Published: 13 October 2015

Reviewing editor: M Dawn Teare, University of Sheffield, United Kingdom

Copyright Henderson et al. This article is distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use and redistribution provided that the original author and source are credited.

Introduction

Preclinical experiments provide evidence of clinical promise, inform trial design, and establish the ethical basis for exposing patients to a new substance. However, preclinical research is plagued by poor design and reporting practices (***van der Worp et al., 2010***; ***Begley, 2013a***; ***Begley and Ioannidis, 2015***). Recent reports also suggest that many effects in preclinical studies fail replication (***Begley and Ellis, 2012***). Drug development efforts grounded on non-reproducible findings expose patients to harmful and inactive agents; they also absorb scarce scientific and human resources, the costs of which are reflected as higher drug prices.

Several studies have evaluated the predictive value of animal models in cancer drug development (***Johnson et al., 2001***; ***Voskoglou-Nomikos et al., 2003***; ***Corpet and Pierre, 2005***). However, few have systematically examined experimental design—as opposed to use of specific models—and its impact on effect sizes across different malignancies (***Amarasingh et al., 2009***; ***Hirst et al., 2013***). A recent systematic review of guidelines for limiting bias in preclinical research design was unable to identify any guidelines in oncology (***Henderson et al., 2013***). Validity threats in preclinical oncology may be particularly important to address in light of the fact that cancer drug development has one of the highest rates of attrition (***Hay et al., 2014***), and oncology drug development commands billions of dollars in funding each year (***Adams and Brantner, 2006***).

In what follows, we conducted a systematic review and meta-analysis of features of design and outcomes for preclinical efficacy studies of the highly successful drug sunitinib. Sunitinib is a multi-targeted tyrosine kinase inhibitor sunitinib (SU11248, Sutent) and is licensed as monotherapy for three

eLife digest Developing a new drug can take years, partly because preclinical research on non-human animals is required before any clinical trials with humans can take place. Nevertheless, only a fraction of cancer drugs that are put into clinical trials after showing promising results in preclinical animal studies end up proving safe and effective in human beings.

Many researchers and commentators have suggested that this high failure rate reflects flaws in the way preclinical studies in cancer are designed and reported. Now, Henderson et al. have looked at all the published animal studies of a cancer drug called sunitinib and asked how well the design of these studies attempted to limit bias and match the clinical scenarios they were intended to represent.

This systematic review and meta-analysis revealed that many common practices, like randomization, were rarely implemented. None of the published studies used 'blinding', whereby information about which animals are receiving the drug and which animals are receiving the control is kept from the experimenter, until after the test; this technique can help prevent any expectations or personal preferences from biasing the results. Furthermore, most tumors were tested in only one model system, namely, mice that had been injected with specific human cancer cells. This makes it difficult to rule out that any anti-cancer activity was in fact unique to that single model.

Henderson et al. went on to find evidence that suggests that the anti-cancer effects of sunitinib might have been overestimated by as much as 45% because those studies that found no or little anti-cancer effect were simply not published. Though it is known that the anti-cancer activity of the drug increases with the dose given in both human beings and animals, an evaluation of the effects of all the published studies combined did not detect such a dose-dependent response.

The poor design and reporting issues identified provide further grounds for concern about the value of many preclinical experiments in cancer. These findings also suggest that there are many opportunities for improving the design and reliability of study reports. Researchers studying certain medical conditions (such as strokes) have already developed, and now routinely implement, a set of standards for the design and reporting of preclinical research. It now appears that the cancer research community should do the same.

DOI: 10.7554/eLife.08351.002

different malignancies (***Chow and Eckhardt, 2007***; ***Raymond et al., 2011***). As it was introduced into clinical development around 2000 and tested against numerous malignancies, sunitinib provided an opportunity to study a large sample of preclinical studies across a broad range of malignancies—including several supporting successful translation trajectories.

Results

Study characteristics

Our screen from database and reference searches captured 74 studies eligible for extraction, corresponding to 332 unique experiments investigating tumor volume response (***Figure 1***, ***Table 1***, ***Table 1—source data 1E***). Effect sizes (standardized mean difference [SMD] using Hedges' g) could not be computed for 174 experiments (52%) due to inadequate reporting (e.g., sample size not provided, effect size reported as a median, lack of error bars, ***Figure 1—figure supplement 1***). Overall, 158 experiments, involving 2716 animals, were eligible for meta-analysis. The overall pooled SMD for all extracted experiments across all malignancies was −1.8 [−2.1, −1.6] (***Figure 2—figure supplement 1***). Mean duration of experiments used in meta-analysis (***Figures 2–4***) was 31 days (±14 days standardized deviation of the mean (SDM)).

Design elements addressing validity threats

Effects in preclinical studies can fail clinical generalization because of bias or random variation (internal validity), a mismatch between experimental operations and the clinical scenario modeled (construct validity), or idiosyncratic causal mediators in an experimental system (external validity) (***Henderson et al., 2013***). We extracted design elements addressing each using consensus design practices identified in a systematic review of validity threats in preclinical research (***Henderson et al., 2013***).

A

Internal Validity Characteristics	Experiments (n=158)
Exact sample size given for all groups	120 (76%)
Performance of sample size calculation	0 (0%)
Randomly allocated to treatment	58 (37%)
Concealed allocation	0 (0%)
Blinded outcome assessment	0 (0%)
Used named inferential statistical test	117 (74%)
Addressed animal flow through experiment	21 (13%)
Evaluated dose-response (≥3 doses)	9 (6%)
Construct Validity Characteristics	
Species	
Mouse	156 (99%)
Rat	2 (1%)
Age	
Pediatric/Juvenile	78 (49%)
Adult	25 (16%)
Aged	0 (0%)
Immune Status	
Immunocompetent	30 (19%)
Immunocompromised	128 (81%)
Sex	
Male	12 (8%)
Female	104 (66%)
Model Type	
Human xenograft	110 (70%)
Allografts	42 (27%)
Genetically engineered	4 (2%)
Other	2 (1%)
Experiment type	
Late-Stage Tumour Model	63 (40%)
Early-Stage Tumour Model	91 (58%)
Evidence for causal mechanism	
Molecular	66 (42%)
Physiological	123 (78%)
Functional/Clinical/Behavioural	0 (0%)

B

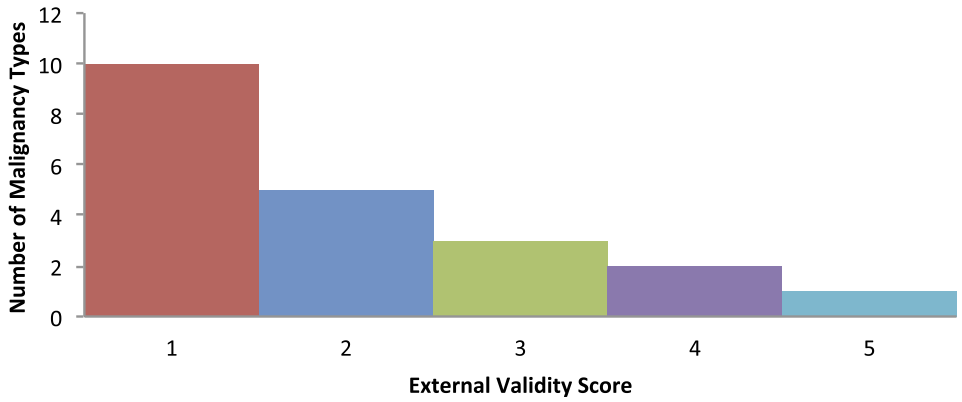

Figure 1. Descriptive analysis of (**A**) internal, construct, and (**B**) external validity design elements. External validity scores were calculated for each malignancy type tested, according to the formula: number species used + number of models used; an extra point was assigned if a malignancy type tested more than one species *and* more than one model.

DOI: 10.7554/eLife.08351.003

The following source data and figure supplement are available for figure 1:

Source data 1. (A) Coding details for IV and CV categories.

DOI: 10.7554/eLife.08351.004

Figure 1. continued on next page

Figure 1. Continued

Figure supplement 1. Descriptive analysis of (**A**) internal, construct, and (**B**) external validity design elements for all experiments (n = 332) extracted for validity data parameters.
DOI: 10.7554/eLife.08351.005

Few studies used practices like blinding or randomization to address internal validity threats (***Figure 1A***). Only 6% of experiments investigated a dose–response relationship (3 or more doses). Concealment of allocation or blinded outcome assessment was never reported in studies that advanced to meta-analysis. It is worth noting that one research group employed concealed allocation and blinded assessment for the many experiments it described (***Maris et al., 2008***). However, statistics were reported in a way that did not align with those we needed to calculate SMD. We found that 58.8% of experiments included active drug comparators, thus, facilitating interpretation of sunitinib activity (however, we note that in some of the experiments, sunitinib was an active comparator in a test of a different drug or drug combination). Construct validity practices can only be meaningfully evaluated against a particular, matched clinical trial. Nevertheless, ***Figure 1A*** shows that experiments predominantly relied on juvenile, female, immunocompromised mouse models, and very few animal efficacy experiments used genetically engineered cancer models (n = 4) or spontaneously arising tumors (n = 0). Malignancies generally scored low (score = 1) for addressing external validity (***Figure 1B***), with breast cancer studies employing the greatest variety of species (n = 2) and models (n = 4).

Implementation of internal validity practices did not show clear relationships with effect sizes (***Figure 3A***). However, sunitinib effect sizes were significantly greater when active drug comparators were present in an experiment compared to when they were not (−2.2 [−2.5, −1.9] vs −1.4 [−1.7, −1.1], p-value <0.001).

Within construct validity, there was a significant difference in pooled effect size between genetically engineered mouse models and human xenograft (p-value <0.0001) and allograft (p-value 0.001) model types (***Figure 3B***). For external validity (***Figure 3C***), malignancies tested in more and diverse experimental systems tended to show less extreme effect sizes (p < 0.001).

Table 1. Demographics of included studies

Study level demographics	Included studies (n = 74)
Conflict of interest	
Declared	19 (26%)
Funding statement*	
Private, for-profit	44 (59%)
Private, not-for-profit	35 (47%)
Public	37 (50%)
Other	2 (3%)
Recommended clinical testing	
Yes	37 (50%)
Publication date	
2003–2006	13 (18%)
2007–2009	17 (23%)
2010–2013	44 (59%)

*Does not sum to 100% as many studies declared more than one funding source.
DOI: 10.7554/eLife.08351.006

Source data 1. (C) Search Strategies. (D) PRISMA Flow Diagram. (E) Demographics of included studies at qualitative level.
DOI: 10.7554/eLife.08351.007

Evidence of publication bias

For the 158 individual experiments, 65.8% showed statistically significant activity at the experiment level ($p < 0.05$, ***Figure 2—figure supplement 1***), with an average sample size of 8.03 animals per treatment arm and 8.39 animals per control arm. Funnel plots for all studies (***Figure 4A***), as well as our renal cell carcinoma (RCC) subset (***Figure 4B***) suggest potential publication bias. Trim and fill analysis suggests an overestimation of effect size of 45% (SMD changed from −1.8 [−2.1, −1.7] to −1.3 [−1.5, −1.0]) across all indications. For high-grade glioma and breast cancer, the overestimation was 11% and 52%, respectively. However, trim and fill analysis suggested excellent symmetry for the RCC subgroup, suggesting coverage of the overall effect size and confidence intervals and not overestimation of effect size.

Preclinical studies and clinical correlates

Every malignancy tested with sunitinib showed statistically significant anti-tumor activity (***Figure 2***). Though we did not perform a systematic review to estimate clinical effect sizes for sunitinib against various malignancies, a perusal of the clinical literature suggests little relationship between pooled effect sizes and demonstrated clinical activity. For instance, sunitinib monotherapy is highly active in RCC patients (***Motzer et al., 2006a***, ***2006b***) and yet showed a relatively small preclinical effect; in contrast, sunitinib monotherapy was inactive against small cell lung cancer in a phase 2 trial (***Han et al., 2013***), but showed relatively large preclinical effects.

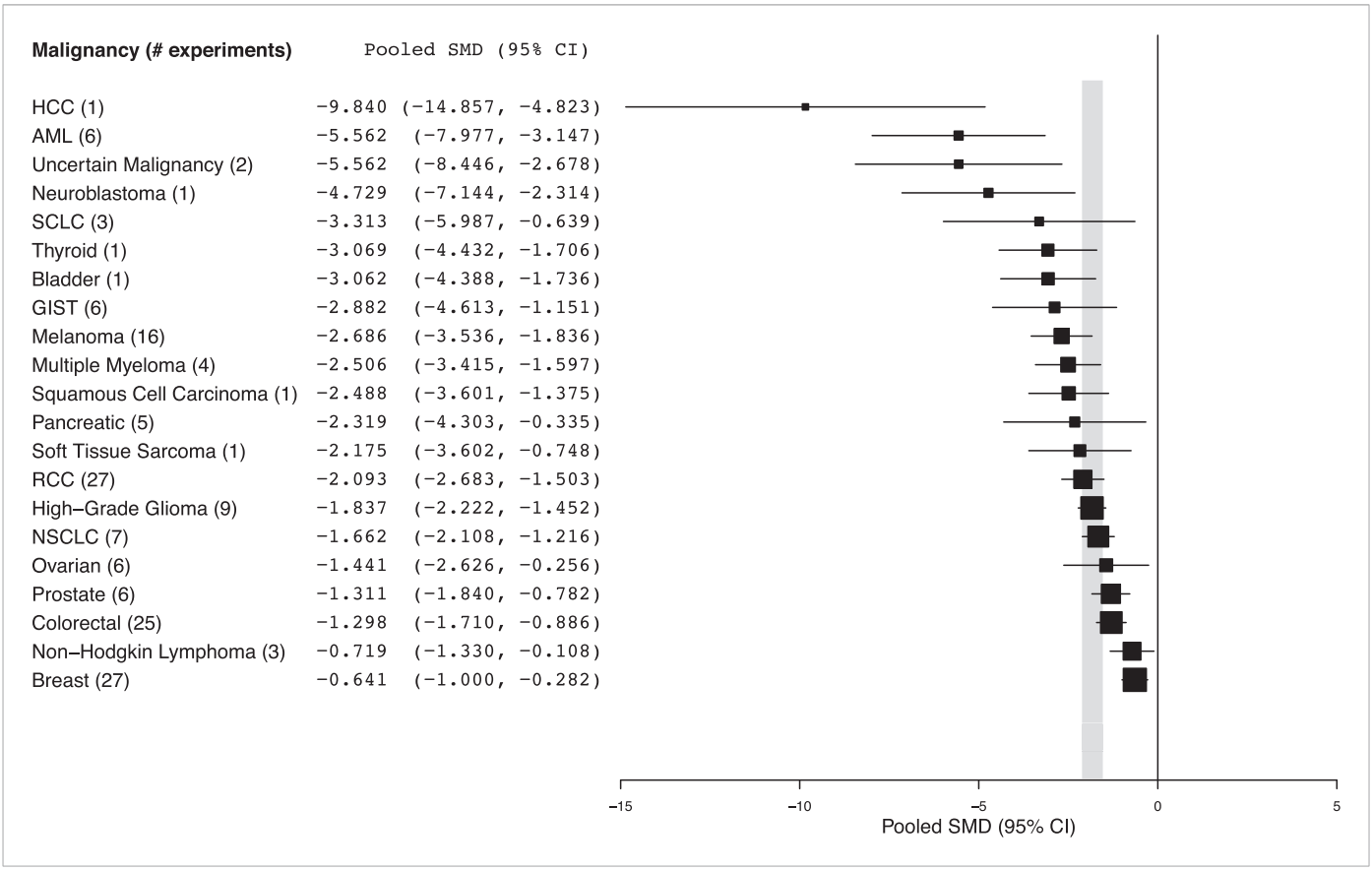

Figure 2. Summary of pooled SMDs for each malignancy type. Shaded region denotes the pooled standardized mean difference (SMD) and 95% confidence interval (CI) (−1.8 [−2.1, −1.6]) for all experiments combined at the last common time point (LCT).

DOI: 10.7554/eLife.08351.008

The following source data and figure supplement are available for figure 2:

Source data 1. (B) Heterogeneity statistics (I^2) for each malignancy sub-group.

DOI: 10.7554/eLife.08351.009

Figure supplement 1. Effect sizes for all included experiments (n = 158).

DOI: 10.7554/eLife.08351.010

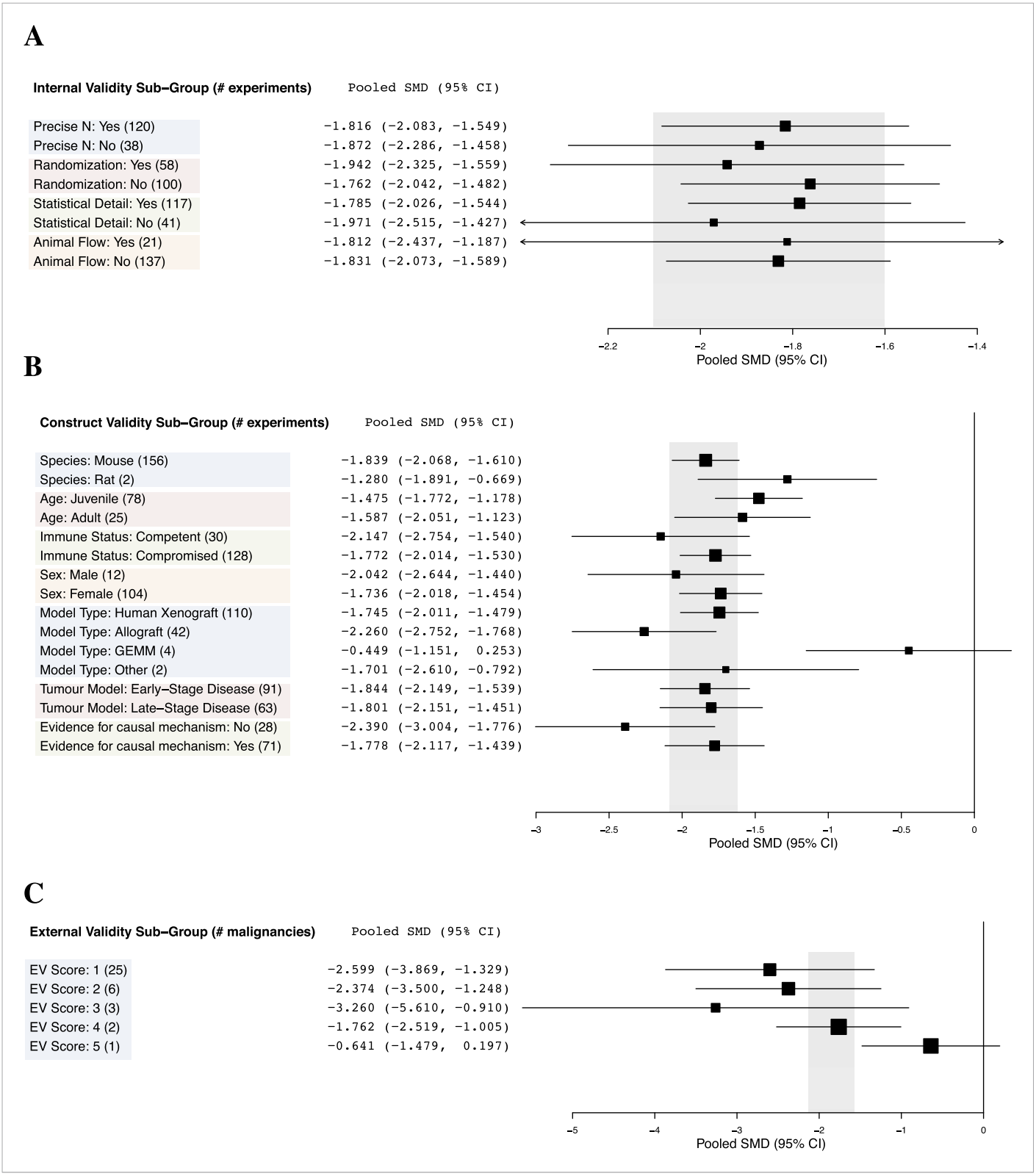

Figure 3. Relationship between study design elements and effect sizes. The shaded region denotes the pooled SMD and 95% CI (−1.8 [−2.1, −1.6]) for all experiments combined at the LCT.

DOI: 10.7554/eLife.08351.011

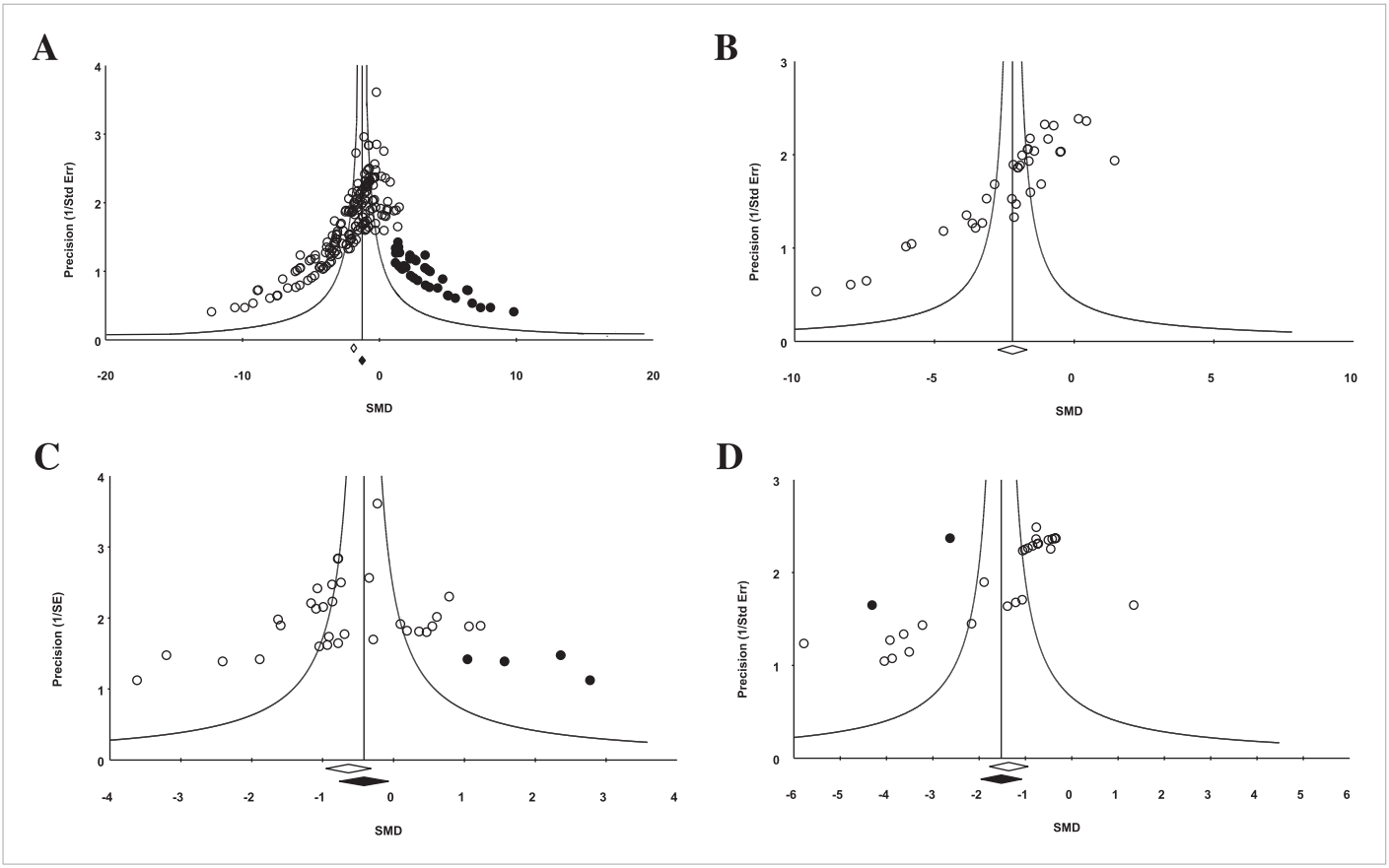

Figure 4. Funnel plot to detect publication bias. Trim and fill analysis was performed on pooled malignancies, as well as the three malignancies with the greatest study volume. (**A**) All experiments for all malignancies (n = 182), (**B**) all experiments within renal cell carcinoma (RCC) (n = 35), (**C**) breast cancer (n = 32), and (**D**) colorectal cancer (n = 29). Time point was the LCT. Open circles denote original data points whereas black circles denote 'filled' experiments. Trim and fill did not produce an estimate in RCC; therefore, no overestimation of effect size could be found.

DOI: 10.7554/eLife.08351.012

Using measured effect sizes at a standardized time point of 14 days after first administration (a different time point than in ***Figures 2–4*** to better align our evaluation of dose–response), we were unable to observe a dose–response relationship over three orders of magnitude (0.2–120 mg/kg/day) for all experiments (***Figure 5A***). We were also unable to detect a dose–response relationship over the full dose range (4–80 mg/kg/day) tested in the RCC subset (***Figure 5B***). The same results were observed when we performed the same analyses using the last time point in common between the experimental and control arms.

Discussion

Preclinical studies serve an important role in formulating clinical hypotheses and justifying the advance of a new drug into clinical testing. Our meta-analysis, which included malignancies that respond to sunitinib in human beings and those that do not, raises several questions about methods and reporting practices in preclinical oncology—at least in the context of one well-established drug.

First, reporting of design elements and data was poor and inconsistent with widely recognized standards for animal studies (***Kilkenny et al., 2010***). Indeed, 98 experiments (30% of qualitative sample) could not be quantitatively analyzed because sample sizes or measures of dispersion were not provided. Experimenters only sporadically addressed major internal validity threats and tended not to test indication-activity in more than one model and species. This finding is consistent with what others have observed in experimental stroke and other research areas (***Macleod et al., 2004***; ***van der Worp et al., 2005***; ***Kilkenny et al., 2009***; ***Glasziou et al., 2014***). Some teams have shown a relationship between failure to address internal validity threats and exaggerated effect size (***Crossley et al., 2008***;

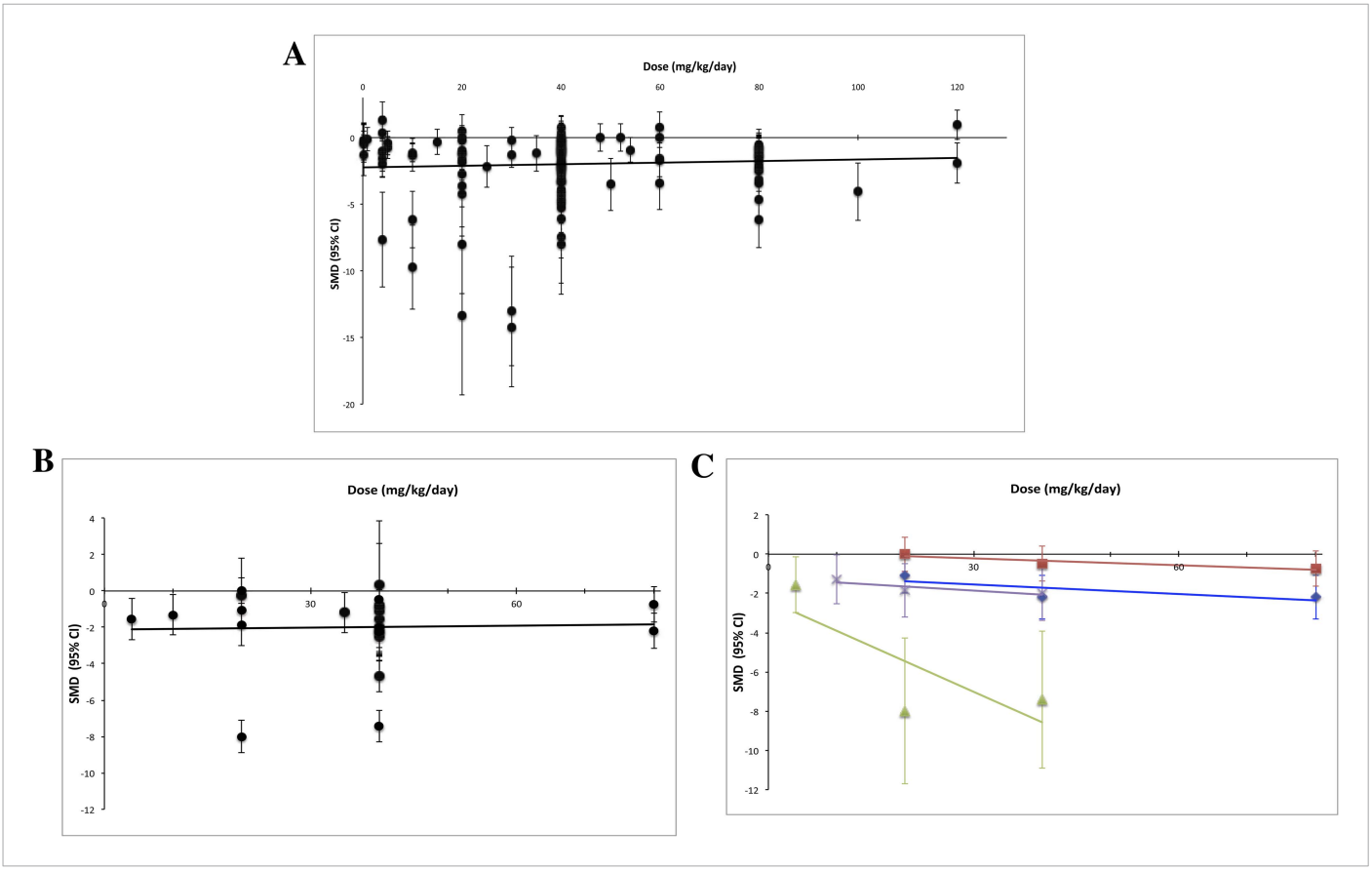

Figure 5. Dose–response curves for sunitinib preclinical studies. Only experiments with a once daily (no breaks) administration schedule were included in both graphs. Effect size data were taken from a standardized time point (14 days after first sunitinib administration). (**A**) Experiments (n = 158) from all malignancies tested failed to show a dose–response relationship. (**B**) A dose–response relationship was not detected for RCC (n = 24). (**C**) Dose–response curves reported in individual studies within the RCC subset showed dose–response patterns (blue diamond = Huang 2010a [n = 3], red square = Huang 2010d [n = 3], green triangle = Ko 2010a [n = 3], purple X = Xin 2009 [n = 3]).

DOI: 10.7554/eLife.08351.013

Rooke et al., 2011); we did not observe a clear relationship. Consistent with what has been reported in stroke (***O'Collins et al., 2006***), our findings suggest that testing in more models tends to produce smaller effect sizes. However, since a larger sample of studies will provide a more precise estimate of effect, we cannot rule out that the trends observed for external validity reflect a regression to the mean.

Second, preclinical studies for sunitinib seem to be prone to publication bias. Notwithstanding limitations on using funnel plots to detect publication bias (***Lau et al., 2006***), our plots were highly asymmetrical. That all malignancy types tested showed statistically significant anti-cancer activity strains credulity. Others have reported that far more animal studies report statistical significance than would be expected (***Wallace et al., 2009***; ***Tsilidis et al., 2013***), and our observations that two thirds of individual studies showed significance extends these observations.

Third, we were unable to detect a meaningful relationship between preclinical effect sizes and known clinical behavior. Although a full analysis correlating trial and preclinical effect sizes will be needed, we did not observe obvious relationships between the two. We also did not detect a dose–response effect over three orders of magnitude even within an indication—RCC—known to respond to sunitinib and even when different time points were used. It is possible that heterogeneity in cell lines or strains may have obscured the effects of dose. For example, experimenters may have delivered higher doses to xenografts known to show slow tumor growth. However, RCC patients—each of whom harbors genetically distinct tumors—show dose–response effects in trials (***Faivre et al., 2006***) and between trials in a meta-analysis (***Houk et al., 2010***). It is also possible that the toxicity of sunitinib may have limited the ability to demonstrate dose response, though this

contradicts demonstration of dose response within studies (**Abrams et al., 2003**; **Amino et al., 2006**; **Ko et al., 2010**). Finally, the tendency for preclinical efficacy studies to report drug dose, but rarely drug exposure (i.e., serum measurement of active drug), further limits the construct validity of these studies (**Peterson and Houghton, 2004**).

One explanation for our findings is that human xenograft models, which dominated our meta-analytic sample, have little predictive value, at least in the context of receptor tyrosine kinase inhibitors. This is a possibility that contradicts other reports (**Kerbel, 2003**; **Voskoglou-Nomikos et al., 2003**). We disfavor this explanation in light of the suggestion of publication bias; also, xenografts should show a dose–response regardless of whether they are useful clinical models. A second explanation is that experimental methods are so varied as to mask real effects. However, we note that the observed patterns on experimental design are based purely on what was reported in 'Materials and methods' section. Third, experiments assessing changes in tumor volume might only be interpretable in the context of other experiments within a preclinical report, such as with mechanistic and pharmacokinetic studies. This explanation is consistent with our observation that studies testing effect along a causal pathway tended to produce smaller effect sizes. A fourth possible explanation for our findings is that the predictive value of a small number of preclinical studies was obscured by inclusion of poorly designed and executed preclinical studies in our meta-analysis. Quantitative analysis of preclinical design factors that confer greater clinical generalizability awaits side-by-side comparison with pooled effects in clinical trials. Finally, it may be that design and reporting practices are so poor in preclinical cancer research as to make interpretation of tumor volume curves useless. Or, non-reporting may be so rampant as to render meta-analysis of preclinical research impossible. If so, this raises very troubling questions for the publication economy of cancer biology: even well-designed and reported studies may be difficult to interpret if their results cannot be compared to and synthesized with other studies.

Our systematic review has several limitations. First, we relied on what authors reported in the published study. It is possible certain experimental practices, like randomization, were used but not reported in methods. Further to this, we relied only on published reports, and restriction of searches to the English language may have excluded some articles. In February of 2012, we filed a Freedom of Information Act request from the Food and Drug Administration (FDA) for additional preclinical data submitted in support of sunitinib's licensure; nearly 4 years later, the request has not been honored. Second, effect sizes were calculated using graph digitizer software from tumor volume curves: minor distortion of effect sizes may have occurred but were likely non-differential between groups. Third, subtle experimental design features—not apparent in 'Materials and methods' sections—may explain our failure to detect a dose–response effect. For instance, few reports provide detailed animal housing and testing conditions, perhaps leading to important inter-laboratory differences in tumor growth. It should also be emphasized that our study was exploratory in nature; findings like ours will need to be confirmed using prespecified protocols. Fourth, our study represents analysis of a single drug, and it may be our findings do not extend beyond receptor tyrosine kinase inhibitors, or sunitinib. However, many of our findings are consistent with those observed in other systematic reviews of preclinical cancer interventions (**Amarasingh et al., 2009**; **Sugar et al., 2012**; **Hirst et al., 2013**). Fifth, our analysis does not directly address many design elements—like duration of experiment or choice of tissue xenograft—that are likely to bear on study validity. Finally, we acknowledge that there may be funding constraints that limit implementation of validity practices described above. We note, nevertheless, that other realms, in particular, neurology, have found ways to make such methods a mainstay.

Numerous commentators have raised concerns about the design and reporting of preclinical cancer research (**Sugar et al., 2012**; **Begley, 2013b**). In one report, only 11% preclinical cancer studies submitted to a major biotechnology company withstood in-house replication (**Begley and Ellis, 2012**). The Center for Open Science and Science Exchange has initiated a project that will attempt to reproduce 50 of the highest impact papers in cancer biology published between 2010 and 2012 (**Morrison, 2014**). In a recent commentary, Smith et al. fault many researchers for performing in vitro preclinical tests using drug levels that are clinically unachievable due to toxicity (**Smith and Houghton, 2013**). Unaddressed preclinical validity threats like this—and the ones documented in our study—encourage futile clinical development trajectories. Many research areas, like stroke, epilepsy, and cardiology, have devised design guidelines aimed at improving the clinical generalizability of preclinical studies (**Fisher et al., 2009**; **Galanopoulou et al., 2012**; **Curtis et al., 2013**; **Pusztai et al., 2013**); and the ARRIVE guidelines (**Kilkenny et al., 2010**) for reporting animal experiments have been

taken up by numerous journals and funding bodies. Our findings provide further impetus for developing and implementing guidelines for the design, reporting, and synthesis of preclinical studies in cancer.

Materials and methods

Literature search

To identify all in vivo animal studies testing the anti-cancer properties of sunitinib ('efficacy studies'), we queried the following databases on 27 February 2012 using a search strategy adapted from ***Hooijmans et al. (2010)*** and ***de Vries et al. (2011)***: Ovid MEDLINE In-Process & Other Non-Indexed Citations and Ovid MEDLINE (dates of coverage from 1948 to 2012), EMBASE Classic and EMBASE database (dates of coverage from 1974 to 2012) and BIOSIS Previews (dates of coverage from 1969 to 2012). Search results were entered into an EndNote library and duplicates were removed. Additional citations were identified during the screening of identified articles. See ***Table 1—source data 1C,D*** for detailed search strategy and PRISMA flow diagram.

Screening was performed at citation level by two reviewers (CF and VCH), and at full-text by one reviewer (VCH). Inclusion criteria were (a) original reports or abstracts, (b) English language, (c) contained at least one experiment measuring disease response in a live, non-human animals, and (d) employed sunitinib in a control, comparator, or experimental context, (e) tested anti-cancer activity. To avoid capturing the same experiment twice, in rare cases where the same experiment was reported in different articles, the most detailed and/or recent publication was included.

Extraction

All included studies were evaluated at the study-level, but only those with eligible experiments (e.g., those evaluating the effect of monotherapy on tumor volume and that were reported with sample sizes and error measurements) were forwarded to experiment-level extractions. We excluded experiments when they had been reported in a previous publication after specifically searching for duplicates during screening and analysis. For each eligible experiment, we extracted experimental design elements derived from a prior systematic review of validity threats in preclinical research (***Henderson et al., 2013***).

Details regarding the coding of internal and construct validity categories are given in ***Figure 1—source data 1A***. To score for external validity, we created an index that summed the number of species and models tested for a given malignancy and awarded an extra point if more than one species *and* model was tested. For example, if experiments within a malignancy tested two species and three different model types, the external validity score would be 4 (1 point for the second species, one point for the second model type, one point for the third model type, and an extra point because more than one model and species were employed).

Our primary outcome was experimental tumor volume and we extracted necessary information (sample size, mean measure of treatment effect, and SDM/SEM) to enable calculation of study and aggregate level effect sizes. Since the units of tumor volume were not always consistent between experiments, we extracted those experiments for which a reasonable proxy of tumor volume could be obtained. These included physical caliper measurements (often reported in mm^3 or cm^3), tumor weights (often reported in mg), optical measurements made from luminescent tumor cell lines (often reported in photons/second), and fold differences in tumor volumes between the control and treatment arms. We extracted experiments of both primary and metastatic tumors, but not experiments where tumor incidence was reported. To account for these different measures of tumor volume, SMDs were calculated using Hedges' g. Hedges' g is a widely accepted standardized measure of effect in meta-analyses where units are not always identical. For experiments where more than one dose of sunitinib was tested against the same control arm, we created a pooled SMD to adjust appropriately for the multiple use of the same control group. Data were extracted at baseline (Day 0 and defined as the first day of drug administration), Day 14 (the closest measured data point to 14 days following first dose), and the last common time point (LCT) between the control group and the treatment group. The LCT was variable between experiments and the last time point for which we could calculate SMD and often represented the point at which the greatest difference was observed between the arms. Data presented graphically were extracted using the graph digitizer software GraphClick (Arizona Software). Extraction was performed by four independent and trained coders

(VCH, ND, AH, and NM) using DistillerSR. There was a 12% double-coding overlap to minimize inter-rater heterogeneity and prevent coder drift. Discrepancies in double coding were reconciled through discussion, and if necessary, by a third coder. The gross agreement rate before reconciliation for all double-coded studies was 83%.

Meta-analysis

Effect sizes were calculated as SMDs using Hedges' g with 95% confidence intervals. Pooled effect sizes were calculated using a random effects model employing the ***DerSimonian and Laird (1986)*** method, in OpenMeta[Analyst] (***Wallace et al., 2009***). We also calculated heterogeneity within each malignancy using I^2 statistics (***Figure 2—source data 1B***). To assess the predictive value of preclinical studies in our sample, we calculated pooled effect sizes for each type of malignancy. Subgroup analyses were performed for each validity element. p-values were calculated by a two-sided independent group T-test. Statistical significance was set at a p-value <0.05; as this was an exploratory study we did not adjust for multiple analyses.

Funnel plots to assess publication bias and Duval and Tweedie's trim and fill estimates were generated using Comprehensive Meta Analyst software (***Dietz et al., 2014***). Funnel plots were created for all experiments in aggregate, and for the three indications for which greater than 20 experiments were analyzable.

Dose–response curves are a widely used tool for testing the strength of causal relationships (***Hill, 1965***), and if preclinical studies indicate real drug-responses, we should be able to detect a dose–response effect across different experiments. Dose–response relationships were found in post-analysis of sunitinib clinical studies in metastatic RCC and Gastrointestinal stromal tumour (GIST) (***Houk et al., 2010***). We tested for all indications in aggregate, as well as for RCC, an indication known to respond to sunitinib in human beings (***Motzer et al., 2006a***, ***2006b***, ***2009***). To eliminate variation at the LCT between treatment and control arms, dose–response curves were created using data from a time point 14 days from the initiation of sunitinib treatment. Experiments with more than one treatment arm were not pooled as in other analyses, but expanded out so that each treatment arm (with it's respective dose) could be plotted properly. As we were unable to find experiments that reported drug exposure (e.g., drug serum levels), we calculated pooled effect sizes in OpenMeta[Analyst] and plotted against dose. To avoid the confounding effect of discontinuous dosing, we included only experiments that used a regular administration schedule without breaks (i.e., sunitinib administered at a defined dose once a day instead of experiments where sunitinib was dosed more irregularly or only once).

As this meta-analysis was exploratory and involved development of methodology, we did not prospectively register a protocol.

Acknowledgements

Dan G Hackam, Jeremy Grimshaw, Malcolm Smith, Elham Sabri, Benjamin Carlisle.

Additional information

Funding

Funder	Grant reference	Author
Canadian Institutes of Health Research (Instituts de recherche en santé du Canada)	EOG 111391	Jonathan Kimmelman

The funder had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

VCH, JK, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; ND, AH, NMK, CAF, Acquisition of data, Drafting or revising the article; DF, Conception and design, Analysis and interpretation of data, Drafting or revising the article

References

Anonymous. 2014. Budgets up at NIH, NCI, and FDA. *Cancer Discovery* **4**:263. doi: 10.1158/2159-8290.CD-NB2014-016.

Abrams TJ, Murray LJ, Pesenti E, Holway VW, Colombo T, Lee LB, Cherrington JM, Pryer NK. 2003. Preclinical evaluation of the tyrosine kinase inhibitor SU11248 as a single agent and in combination with 'standard of care' therapeutic agents for the treatment of breast cancer. *Molecular Cancer Therapeutics* **2**:1011–1021.

Adams CP, Brantner VV. 2006. Estimating the cost of new drug development: is it really 802 million dollars? *Health Affairs* **25**:420–428. doi: 10.1377/hlthaff.25.2.420.

Amarasingh S, Macleod MR, Whittle IR. 2009. 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. *Journal of Neuro-Oncology* **91**:117–125. doi: 10.1007/s11060-008-9697-z.

Amino N, Ideyama Y, Yamano M, Kuromitsu S, Tajinda K, Samizu K, Hisamichi H, Matsuhisa A, Shirasuna K, Kudoh M, Shibasaki M. 2006. YM-359445, an orally bioavailable vascular endothelial growth factor receptor-2 tyrosine kinase inhibitor, has highly potent antitumor activity against established tumors. *Clinical Cancer Research* **12**:1630–1638. doi: 10.1158/1078-0432.CCR-05-2028.

Begley CG. 2013a. Six red flags for suspect work. *Nature* **497**:433–434. doi: 10.1038/497433a.

Begley CG. 2013b. An unappreciated challenge to oncology drug discovery: pitfalls in preclinical research. *American Society of Clinical Oncology Educational Book* **2013**:466–468. doi: 10.1200/EdBook_AM.2013.33.466.

Begley CG, Ellis LM. 2012. Drug development: raise standards for preclinical cancer research. *Nature* **483**:531–533. doi: 10.1038/483531a.

Begley CG, Ioannidis JP. 2015. Reproducibility in science: improving the standard for basic and preclinical research. *Circulation Research* **116**:116–126. doi: 10.1161/CIRCRESAHA.114.303819.

Chow LQ, Eckhardt SG. 2007. Sunitinib: from rational design to clinical efficacy. *Journal of Clinical Oncology* **25**:884–896. doi: 10.1200/JCO.2006.06.3602.

Corpet DE, Pierre F. 2005. How good are rodent models of carcinogenesis in predicting efficacy in humans? A systematic review and meta-analysis of colon chemoprevention in rats, mice and men. *European Journal of Cancer* **41**:1911–1922. doi: 10.1016/j.ejca.2005.06.006.

Crossley NA, Sena E, Goehler J, Horn J, van der Worp B, Bath PM, Macleod M, Dirnagl U. 2008. Empirical evidence of bias in the design of experimental stroke studies: a metaepidemiologic approach. *Stroke* **39**:929–934. doi: 10.1161/STROKEAHA.107.498725.

Curtis MJ, Hancox JC, Farkas A, Wainwright CL, Stables CL, Saint DA, Clements-Jewery H, Lambiase PD, Billman GE, Janse MJ, Pugsley MK, Ng GA, Roden DM, Camm AJ, Walker MJ. 2013. The Lambeth Conventions (II): guidelines for the study of animal and human ventricular and supraventricular arrhythmias. *Pharmacology & Therapeutics* **139**:213–248. doi: 10.1016/j.pharmthera.2013.04.008.

de Vries RB, Hooijmans CR, Tillema A, Leenaars M, Ritskes-Hoitinga M. 2011. A search filter for increasing the retrieval of animal studies in Embase. *Laboratory Animals* **45**:268–270. doi: 10.1258/la.2011.011056.

DerSimonian R, Laird N. 1986. Meta-analysis in clinical trials. *Contemporary Clinical Trials* **7**:177–188. doi: 10.1016/0197-2456(86)90046-2.

Dietz G, Dahabreh IJ, Gurevitch J, Lajeunesse MJ, Schmid CH, Trikalinos TA, Wallace BC. 2014. *OpenMEE: software for ecological and evolutionary meta-analysis*.

Faivre S, Delbaldo C, Vera K, Robert C, Lozahic S, Lassau N, Bello C, Deprimo S, Brega N, Massimini G, Armand JP, Scigalla P, Raymond E. 2006. Safety, pharmacokinetic, and antitumor activity of SU11248, a novel oral multitarget tyrosine kinase inhibitor, in patients with cancer. *Journal of Clinical Oncology* **24**:25–35. doi: 10.1200/JCO.2005.02.2194.

Fisher M, Feuerstein G, Howells DW, Hurn PD, Kent TA, Savitz SI, Lo EH, STAIR Group. 2009. Update of the stroke therapy academic industry roundtable preclinical recommendations. *Stroke* **40**:2244–2250. doi: 10.1161/STROKEAHA.108.541128.

Galanopoulou AS, Buckmaster PS, Staley KJ, Moshé SL, Perucca E, Engel J Jr, Löscher W, Noebels JL, Pitkänen A, Stables J, White HS, O'Brien TJ, Simonato M, American Epilepsy Society Basic Science Committee And The International League Against Epilepsy Working Group On Recommendations For Preclinical Epilepsy Drug Discovery. 2012. Identification of new epilepsy treatments: issues in preclinical methodology. *Epilepsia* **53**:571–582. doi: 10.1111/j.1528-1167.2011.03391.x.

Glasziou P, Altman DG, Bossuyt P, Boutron I, Clarke M, Julious S, Michie S, Moher D, Wager E. 2014. Reducing waste from incomplete or unusable reports of biomedical research. *Lancet* **383**:267–276. doi: 10.1016/S0140-6736(13)62228-X.

Han JY, Kim HY, Lim KY, Han JH, Lee YJ, Kwak MH, Kim HJ, Yun T, Kim HT, Lee JS. 2013. A phase II study of sunitinib in patients with relapsed or refractory small cell lung cancer. *Lung Cancer* **79**:137–142. doi: 10.1016/j.lungcan.2012.09.019.

Hay M, Thomas DW, Craighead JL, Economides C, Rosenthal J. 2014. Clinical development success rates for investigational drugs. *Nature Biotechnology* **32**:40–51. doi: 10.1038/nbt.2786.

Henderson VC, Kimmelman J, Fergusson D, Grimshaw JM, Hackam DG. 2013. Threats to validity in the design and conduct of preclinical efficacy studies: a systematic review of guidelines for in vivo animal experiments. *PLOS Medicine* **10**:e1001489. doi: 10.1371/journal.pmed.1001489.

Hill AB. 1965. The environment and disease: association or causation? *Proceedings of the Royal Society of Medicine* **58**:295–300.

Hirst TC, Vesterinen HM, Sena ES, Egan KJ, Macleod MR, Whittle IR. 2013. Systematic review and meta-analysis of temozolomide in animal models of glioma: was clinical efficacy predicted? *British Journal of Cancer* **108**:64–71. doi: 10.1038/bjc.2012.504.

Hooijmans CR, Tillema A, Leenaars M, Ritskes-Hoitinga M. 2010. Enhancing search efficiency by means of a search filter for finding all studies on animal experimentation in PubMed. *Laboratory Animals* **44**:170–175. doi: 10.1258/la.2010.009117.

Houk BE, Bello CL, Poland B, Rosen LS, Demetri GD, Motzer RJ. 2010. Relationship between exposure to sunitinib and efficacy and tolerability endpoints in patients with cancer: results of a pharmacokinetic/pharmacodynamic meta-analysis. *Cancer Chemotherapy and Pharmacology* **66**:357–371. doi: 10.1007/s00280-009-1170-y.

Johnson JI, Decker S, Zaharevitz D, Rubinstein LV, Venditti JM, Schepartz S, Kalyandrug S, Christian M, Arbuck S, Hollingshead M, Sausville EA. 2001. Relationships between drug activity in NCI preclinical in vitro and in vivo models and early clinical trials. *British Journal of Cancer* **84**:1424–1431. doi: 10.1054/bjoc.2001.1796.

Kerbel RS. 2003. Human tumor xenografts as predictive preclinical models for anticancer drug activity in humans: better than commonly perceived-but they can be improved. *Cancer Biology & Therapy* **2**(4 Suppl 1):S134–S139.

Kilkenny C, Browne WJ, Cuthill IC, Emerson M, Altman DG. 2010. Improving bioscience research reporting: the ARRIVE guidelines for reporting animal research. *PLOS Biology* **8**:e1000412. doi: 10.1371/journal.pbio.1000412.

Kilkenny C, Parsons N, Kadyszewski E, Festing MF, Cuthill IC, Fry D, Hutton J, Altman DG. 2009. Survey of the quality of experimental design, statistical analysis and reporting of research using animals. *PLOS ONE* **4**:e7824. doi: 10.1371/journal.pone.0007824.

Ko JS, Rayman P, Ireland J, Swaidani S, Li G, Bunting KD, Rini B, Finke JH, Cohen PA. 2010. Direct and differential suppression of myeloid-derived suppressor cell subsets by sunitinib is compartmentally constrained. *Cancer Research* **70**:3526–3536. doi: 10.1158/0008-5472.CAN-09-3278.

Lau J, Ioannidis JP, Terrin N, Schmid CH, Olkin I. 2006. The case of the misleading funnel plot. *BMJ* **333**:597–600. doi: 10.1136/bmj.333.7568.597.

Macleod MR, O'Collins T, Howells DW, Donnan GA. 2004. Pooling of animal experimental data reveals influence of study design and publication bias. *Stroke* **35**:1203–1208. doi: 10.1161/01.STR.0000125719.25853.20.

Maris JM, Courtright J, Houghton PJ, Morton CL, Kolb EA, Lock R, Tajbakhsh M, Reynolds CP, Keir ST, Wu J, Smith MA. 2008. Initial testing (stage 1) of sunitinib by the pediatric preclinical testing program. *Pediatric Blood & Cancer* **51**:42–48. doi: 10.1002/pbc.21535.

Morrison SJ. 2014. Time to do something about reproducibility. *eLife* **3**:e03981. doi: 10.7554/eLife.03981.

Motzer RJ, Hutson TE, Tomczak P, Michaelson MD, Bukowski RM, Oudard S, Negrier S, Szczylik C, Pili R, Bjarnason GA, Garcia-del-Muro X, Sosman JA, Solska E, Wilding G, Thompson JA, Kim ST, Chen I, Huang X, Figlin RA. 2009. Overall survival and updated results for sunitinib compared with interferon alfa in patients with metastatic renal cell carcinoma. *Journal of Clinical Oncology* **27**:3584–3590. doi: 10.1200/JCO.2008.20.1293.

Motzer RJ, Michaelson MD, Redman BG, Hudes GR, Wilding G, Figlin RA, Ginsberg MS, Kim ST, Baum CM, DePrimo SE, Li JZ, Bello CL, Theuer CP, George DJ, Rini BI. 2006a. Activity of SU11248, a multitargeted inhibitor of vascular endothelial growth factor receptor and platelet-derived growth factor receptor, in patients with metastatic renal cell carcinoma. *Journal of Clinical Oncology* **24**:16–24. doi: 10.1200/JCO.2005.02.2574.

Motzer RJ, Rini BI, Bukowski RM, Curti BD, George DJ, Hudes GR, Redman BG, Margolin KA, Merchan JR, Wilding G, Ginsberg MS, Bacik J, Kim ST, Baum CM, Michaelson MD. 2006b. Sunitinib in patients with metastatic renal cell carcinoma. *JAMA* **295**:2516–2524. doi: 10.1001/jama.295.21.2516.

O'Collins VE, Macleod MR, Donnan GA, Horky LL, van der Worp BH, Howells DW. 2006. 1,026 experimental treatments in acute stroke. *Annals of Neurology* **59**:467–477. doi: 10.1002/ana.20741.

Peterson JK, Houghton PJ. 2004. Integrating pharmacology and in vivo cancer models in preclinical and clinical drug development. *European Journal of Cancer* **40**:837–844. doi: 10.1016/j.ejca.2004.01.003.

Pusztai L, Hatzis C, Andre F. 2013. Reproducibility of research and preclinical validation: problems and solutions. *Nature Reviews Clinical Oncology* **10**:720–724. doi: 10.1038/nrclinonc.2013.171.

Raymond E, Dahan L, Raoul JL, Bang YJ, Borbath I, Lombard-Bohas C, Valle J, Metrakos P, Smith D, Vinik A, Chen JS, Hörsch D, Hammel P, Wiedenmann B, Van Cutsem E, Patyna S, Lu DR, Blanckmeister C, Chao R, Ruszniewski P. 2011. Sunitinib malate for the treatment of pancreatic neuroendocrine tumors. *The New England Journal of Medicine* **364**:501–513. doi: 10.1056/NEJMoa1003825.

Rooke ED, Vesterinen HM, Sena ES, Egan KJ, Macleod MR. 2011. Dopamine agonists in animal models of Parkinson's disease: a systematic review and meta-analysis. *Parkinsonism & Related Disorders* **17**:313–320. doi: 10.1016/j.parkreldis.2011.02.010.

Smith MA, Houghton P. 2013. A proposal regarding reporting of in vitro testing results. *Clinical Cancer Research* **19**:2828–2833. doi: 10.1158/1078-0432.CCR-13-0043.

Sugar E, Pascoe AJ, Azad N. 2012. Reporting of preclinical tumor-graft cancer therapeutic studies. *Cancer Biology & Therapy* **13**:1262–1268. doi: 10.4161/cbt.21782.

Tsilidis KK, Panagiotou OA, Sena ES, Aretouli E, Evangelou E, Howells DW, Al-Shahi Salman R, Macleod MR, Ioannidis JP. 2013. Evaluation of excess significance bias in animal studies of neurological diseases. *PLOS Biology* **11**:e1001609. doi: 10.1371/journal.pbio.1001609.

van der Worp HB, de Haan P, Morrema E, Kalkman CJ. 2005. Methodological quality of animal studies on neuroprotection in focal cerebral ischaemia. *Journal of Neurology* **252**:1108–1114. doi: 10.1007/s00415-005-0802-3.

van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, Macleod MR. 2010. Can animal models of disease reliably inform human studies? *PLOS Medicine* **7**:e1000245. doi: 10.1371/journal.pmed.1000245.

Voskoglou-Nomikos T, Pater JL, Seymour L. 2003. Clinical predictive value of the in vitro cell line, human xenograft, and mouse allograft preclinical cancer models. *Clinical Cancer Research* **9**:4227–4239.

Wallace BC, Schmid CH, Lau J, Trikalinos TA. 2009. Meta-analyst: software for meta-analysis of binary, continuous and diagnostic data. *BMC Medical Research Methodology* **9**:80. doi: 10.1186/1471-2288-9-80.