Tenorio-Pedraza *et al.*
Translational Medicine Communications (2023) 8:17
https://doi.org/10.1186/s41231-023-00151-x

Translational Medicine Communications

RESEARCH

Open Access

Meta-analysis of preclinical measures of efficacy in immune checkpoint blockade therapies and comparison to clinical efficacy estimates

Juan Miguel Tenorio-Pedraza[1*], Jörg Lippert[2], Rolf Burghaus[2,3] and Christian Scheerans[2]

Abstract

Background Despite the successes of checkpoint inhibitors targeting T-cell receptors, clinical efficacy is highly cancer-dependent and subject to high inter-individual variability in treatment outcome. The ability to predict the clinical success in different cancer indications is therefore an important capability for successful clinical development. In this meta-analysis, the main goal was to identify factors that modified the clinical efficacy estimates of checkpoint blockade therapies derived from preclinical animal data to improve the robustness and reliability of such estimates.

Methods To this end, animal studies testing checkpoint inhibitors (anti-PD-1, anti-PD-L1, anti-CTLA-4) were identified in PubMed ranging from 1.01.2000 to 31.12.2018. The eligibility criteria included the reporting of the Kaplan–Meier estimates of survival and the number of mice used in each experiment. A mixed-effects model was fitted to the preclinical and clinical data separately to determine potential sources of bias and heterogeneity between studies.

Results A total of 160 preclinical studies comprising 13,811 mice were selected, from which the hazard ratio (HR) and the median survival ratio (MSR) were calculated. Similarly, clinical Phase III studies of checkpoint inhibitors were identified in PubMed and the ClinicalTrials.gov database ranging from 1.01.2010 to 31.12.2020. This resulted in 62 clinical studies representing 43,135 patients subjected to 8 therapies from which overall survival (OS) and progression-free survival (PFS) hazard ratios were obtained. Using a mixed-effects model, different factors were tested to identify sources of variability between estimates. In the preclinical data, the tumor cell line and individual study were the main factors explaining the heterogeneity. In the clinical setting, the cancer type was influential to the inter-study variability. When using the preclinical estimates to predict clinical estimates, the cancer-type specific estimates of treatment effect using the MSRs better approximated the observed clinical estimates than the HR-derived predictions.

Conclusions This has strong implications on the design of ICB preclinical studies with respect to sample size determination, selection of cancer cell lines and labs to run the experiments and the choice of efficacy measure.

*Correspondence:
Juan Miguel Tenorio-Pedraza
jm.tenorio.p@gmail.com
[1] JRC for Computational Biomedicine, RWTH Aachen University, Pauwelsstraße 19, 52074 Aachen, Germany
[2] R&D Pharmaceuticals - Pharmacometrics / Modeling & Simulation, Bayer AG, Leverkusen, Germany
[3] R&D Pharmaceuticals - Systems Pharmacology & Medicine, Bayer AG, Leverkusen, Germany

Background

Animal models for preclinical signaling of drug activity have been a cornerstone of drug development in oncology for decades. More recently, immune checkpoint blockade (ICB) has proven to be more translationally efficient than other treatments in previously resistant types of cancer such as melanoma and lung cancer [1]. ICB studies use syngeneic mouse models (SyMM) with intact immune

systems in contrast to other cancer therapeutics such as chemotherapy, which use xenografts in immunocompromised mice [2, 3]. However, not all checkpoint inhibitors have shown clinical benefit, and despite promising results in preclinical studies, this translational deficit is attributed to faults in experimental design addressing the internal and external validity [4–6]. Internal validity assures that any potential bias can be addressed via the experimental design, whereas external validity maximizes the potential to extrapolate the results from one set of experimental conditions to more clinically relevant ones [5, 6]. These two conditions are threatened by multiple flaws of experimental design, which have been summarized and reviewed in a meta-analysis by Henderson et al. [4] and are presented in greater detail by Worp et al. [6].

The internal validity in preclinical studies is often jeopardized by the failure to implement standard procedures used in clinical trials, such as randomization to treatment, blinded assessment of outcome, and sample size determination. This leads to a potentially significant effect or an overestimation thereof due to the presence of bias. In a meta-analysis of preclinical studies of sunitinib in xenografts, some of these measures were not implemented, which contributed to the over-estimation of treatment effects [7]. Other meta-analyses of preclinical experiments also found this relationship between internal validity and treatment effect overestimation [8, 9]. A different type of bias related to the reporting of the results occurs when the study results are negative, i.e., when no significant effect is detected. Thus, the literature is filled with positive preclinical results, which effectively leads to an overestimation of efficacy [10]. Thus far, neither the internal nor the potential publication bias in preclinical studies of ICB have been characterized and hence it is difficult to determine whether the preclinical efficacy could be overstated.

On the other hand, the external validity characterizes the representation of key aspects of human disease and clinical trials in preclinical experiments, and thus ensuring the applicability of results from the latter setting to the former [6]. These include, e.g., the matching of the clinical population age, sex and baseline state, the use of multiple tumor models to determine the therapeutic range, multiple species testing to account for differences in physiology and immunology as well as independent replication by other labs. However, preclinical studies have multiple shortcomings in the implementation of measures that maximize their translatability [5]. Particularly, the acknowledgement of potentially meaningful sources of variation, such as baseline state (i.e., age, sex, tumor size, inclusion and exclusion criteria), or the laboratory the experiment was performed in, is missing in most studies [11, 12]. Furthermore, there is considerable heterogeneity in responses to ICB treatments across SyMM, which might make some model-dependent effects irrelevant for translational purposes [13]. However, the potential effect of these sources of variability on the interpretation of preclinical readouts and their relevance to clinical outcomes has not been addressed.

In this work, the published preclinical experiments of ICB treatments as monotherapy and in combination with chemo- and radiotherapy were gathered and analyzed to determine their potential biases and unaddressed sources of heterogeneity that might influence translational efficacy in the clinic. Two efficacy measures, the hazard ratio (HR) and the median survival ratio (MSR), were the variables of interest from these studies. The potential internal bias and clinical extrapolation threats in these estimates were investigated to give a qualitative assessment on the implementation of measures that maximize the studies' internal and external validity. The effect of adherence to these measures was quantitatively derived from the comparison of efficacy measures between studies with and without these measures. The publication bias and its potential effect on the efficacy estimates was also addressed. Furthermore, the effect of multiple experimental design variables on the heterogeneity between studies was analyzed to assess which contributed the most to this variability. Finally, the preclinical survival estimates were compared to clinical ones to determine whether there is any correspondence between the two sets of results.

Methods

Literature search and study selection

A search was conducted for studies that fit the following criteria: 1) preclinical mouse studies of ICB, 2) report of survival outcomes via survival curve estimates, 3) report of group sizes, and 4) report of control and monotherapy groups. To collect the preclinical survival studies, a query was entered into PubMed using the following terms: checkpoint blockade OR checkpoint inhibitor OR CD279 OR CD274 OR CD152 OR CTLA-4 OR PD-1 OR PD1 OR PD-L1 OR PD-L1 AND animal model AND survival AND cancer NOT prophylactic NOT Clinical Study [Publication Type] NOT Review [Publication Type]. The time of publication was selected from 01.01.1997 – 31.12.2018. The PRISMA flow diagram for the preclinical studies is shown in Fig. 1a.

For the clinical studies, a search was performed to find studies that met the following criteria: 1) clinical phase 3 studies of ICB in cancer patients, 2) report OS or PFS HRs compared to either the placebo group or the standard of care group for monotherapies, or the monotherapy group for combination therapy, and 3) report of a measure

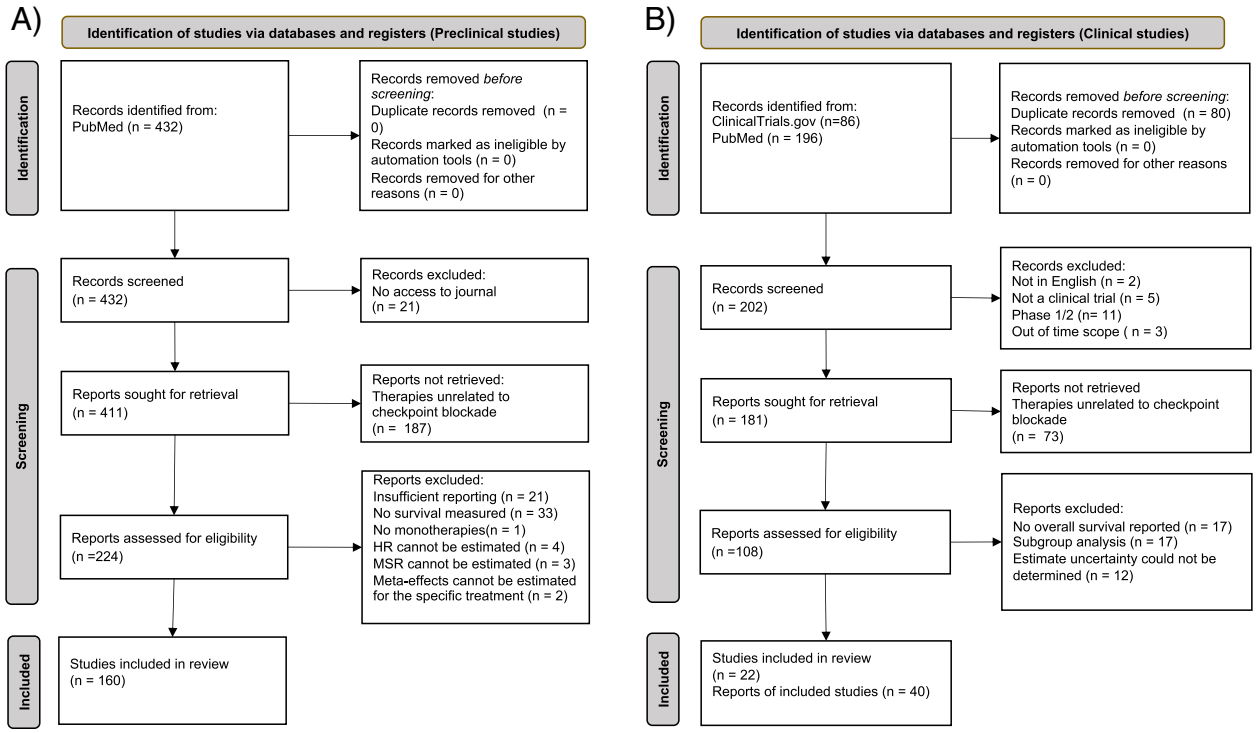

Fig. 1 PRISMA flow diagram for the included studies. (**A**) PRISMA flow diagram for preclinical studies; (**B**) PRISMA flow diagram for clinical studies

of uncertainty of the HRs. To collect the studies, a query was entered into PubMed using the following criteria: ((immune checkpoint inhibitors[Pharmacological Action]) AND ((anti-PD-1) OR (anti-CTLA-4) OR (anti-PD-L1) OR (CD279) OR (CD152) OR (CD274) OR (ipilimumab) OR (tremelimumab) OR (nivolumab) OR (pembrolizumab) OR (avelumab) OR (atezolizumab) OR (durvalumab) OR (PD-L1) OR (PD-1) OR (CTLA-4)) AND ((randomized controlled trial[Publication Type]) OR ((randomized[Title/Abstract]) AND (controlled[Title/Abstract]) AND (trial[Title/Abstract])))) AND ((phase 3[Title/Abstract]) OR (phase III[Title/Abstract])) AND (neoplasms[MeSH Major Topic])) AND (("2000/01/01"[Date—Publication]: "2020/12/31"[Date—Publication])). The time of publication was selected from 01.01.2000 – 31.12.2020. A total of 196 articles were found. These were filtered down to 40 publications that reported the HR of overall survival (OS) and progression free survival (PFS) along with its uncertainty estimate. These studies were complimented a query on the US National Clinical Trial database for phase III clinical trials with publications and results using the following criteria: Studies With Results | Interventional Studies | Cancer | "Anti-CTLA-4" OR "anti-PD-1" OR "anti-PD-L1" OR "ipilimumab" OR "tremelimumab" OR "nivolumab" OR "pembrolizumab" OR "avelumab" OR "atezolizumab" OR "durvalumab" OR "PD-1" OR "PD-L1" OR "CTLA-4" OR "CD279" OR "CD274" OR "CD152" | Phase 3 | Results first posted from 01/01/2000 to 12/31/2020. This resulted in 86 studies, which after removing duplicates, screening, and assessing for eligibility resulted in 22 studies included in the meta-analysis. The full clinical search strategy is depicted in the PRISMA flow diagram in Fig. 1b.

The retrieved studies were independently screened by two reviewers to eliminate articles that did not meet the inclusion criteria. For the qualitative assessment of internal and external validity threats to preclinical studies, all experiments were used. For the meta-analysis, publication bias assessment and heterogeneity analysis, experiments performed in genetically engineered mouse models (GEMM) were excluded. For the comparison to clinical estimates, only those experiments with a similar therapy in the clinic were used (i.e., anti-CTLA-4, anti-PD-1, anti-PD-L1, and two-way combinations with chemotherapy). In the case of clinical studies, all included studies were used for the analysis of potential publication bias, heterogeneity, and final meta-analysis to determine effect sizes.

Data collection

The data was collected by two reviewers independently, any discrepancies were resolved by discussion to reach a consensus. From each preclinical study, survival curves from the control, monotherapy, and combination

therapy groups were all digitized. In some cases, the control group was used multiple times in a single study for comparison to the treatment groups. Hence, only one curve of the control setting was extracted. In other cases, there are multiple control groups throughout the experiments, or the experiments are repeated resulting in a survival curve that is visibly different between the same experimental conditions. In such cases, their survival curves were also extracted. When an efficacy measure such as the HR or median survival is reported, these were also collected to compare to the calculated measures. Furthermore, several items from the design elements of preclinical studies were extracted, according to the recommendations in a previous systematic review of preclinical research guidelines [4]. At the study level, authors names, first author's institutional affiliation, laboratory where experiments took place, and the assessment for translational applications were extracted. At the level of experimental design to minimize bias, the use of randomization to allocate animals to treatment groups, blinded assessment and group size calculation based on statistical power considerations were taken into consideration. Other experimental design elements pertaining to the specifics of the experimental implementation such as cell line, initial tumor cell inoculum, site of inoculation, treatment start day, dose, dose schedule, and route of administration were also collected.

For clinical datasets, the outcome variables collected were the OS HR as well as the PFS HR along with their confidence intervals. Additional variables related to the therapies included drug name, therapy combinations, cancer indication, dose, dosing frequency, control group, and Clinical Trials.gov identifier.

All preclinical datasets were extracted from the figures reported in either the main text or the Supplementary files using the WebPlotDigitizer app for Mac version 4–2. The.tar files documenting alignment of the data extracted with the original figures, as well as the.csv files of the extracted data are available upon request. The final data table with the efficacy measures and experimental design variables is available in the GitHub repository (Survival Meta-Analysis).

Data processing

The preclinical survival curves were back transformed into survival times by taking the unique survival estimates from the survival curves and multiplying by the number of mice in each experiment. The output of this calculation is the size of the at-risk population at each unique event time point (n_{it}) and subtracting n_{it-1} from n_{it}, the number of deaths at time t (d_t) are obtained. The survival times were then used to estimate the log median survival ratio (MSR) and log hazard ratio (HR) between treatment and control groups for each experiment. In the combination therapy, the control group was taken to be the least effective of the single therapies to mimic the clinical scenario where combination therapies are compared to the monotherapy group. Furthermore, in some studies the start time of the survival curve denoted the start of treatment and not the implantation of the tumor in the host, therefore the survival times were shifted to reflect this lag, and in all studies, the time 0 was defined as the time of tumor implantation.

From each preclinical experiment, two efficacy measures were calculated, the Hazard Ratio (HR) and the Median Survival Ratio (MSR) [14] (See Supplementary Information). In some experiments the median survival was not reached. This occurred in 79 experiments (12%), of which 33 (5%) correspond to the therapies of interest in this study (namely, anti-CTLA-4, anti-PD-1, anti-PD-L1 and their two-way combination with each other and with chemotherapy). For these cases, a predictive model was employed to impute this data, using the treatment, the control median survival, and the initial number of tumor cells as predictors (See Supplementary Information for more details). All analyses run using the HRs were also performed with the MSRs, with the number of mice in treatment and control groups serving as an approximation to the MSRs' uncertainty [14]. For quality control of the data gathering and processing, the calculated efficacy measures were compared against the reported ones when available (Supplementary Fig. 1).

Data analysis

The collected preclinical studies were subjected to a qualitative analysis of internal and external validity threats to determine the risk of bias using Henderson's et al. guidelines [4]. All preclinical studies that reported the evaluating criteria were considered for this analysis. A meta-analysis was performed to determine the effect measures summaries (mean differences) for each outcome (HR and MSR) in each therapy. All preclinical studies that had more than two experiments for the same therapy were considered for this synthesis. Then, for each therapy, a trim-and-fill analysis was realized to determine whether there were any reporting biases. Only the experiments with the anti-CTLA-4, anti-PD-1, andi-PD-L1, their two-way combinations and combination with chemotherapy were eligible for this analysis. To identify any potential factors systematically influencing the efficacy readouts, a heterogeneity assessment was undertaken using meta-regression. All treatments that had more than five experiments were used to have sufficient degrees of freedom given the number of free parameters to estimate. To evaluate the influence of the identified modifying factors, a sensitivity analysis was performed by

simulating different experimental designs. The estimates for the eligible therapies were used (anti-CTLA-4, anti-PD-1, andi-PD-L1, their two-way combinations and combination with chemotherapy). The preclinical estimates were compared to clinical ones in two manners: 1) using treatment-wise estimates of efficacy and 2) using treatment-wise and cancer-specific predictions of efficacy. For this, only the estimates using a similar therapy to the clinical trials were used. In the case of clinical studies, all included studies were used for the analysis of potential publication bias, heterogeneity, and final meta-analysis to determine effect sizes.

Meta-analysis

For each preclinical and clinical measure of efficacy, a univariate random effects linear model was fitted to estimate the overall efficacy and its uncertainty in every therapy (further details in Supplementary Information). The heterogeneity between studies was assessed by meta-regression using different experimental design and construct variables as potential explanatory variables [15]. The τ^2 estimate of residual heterogeneity was derived using the restricted-maximum likelihood method [16]. The models were compared based on their estimate of the R^2, i.e. the proportion of heterogeneity captured by the model [14]. The *p*-values of the Cochran's Q-test for heterogeneity were calculated to determine whether there was any significant residual variability between studies after fitting the model [17]. The Akaike Information Criterion (AIC) was used to compare model fits and help guide model selection. All the available preclinical data after extraction was used for this analysis.

To investigate the potential publication bias, two methods were employed: the trim-and-fill analysis and the p-curve analysis. For the former, the dataset was partitioned based on treatment and for each subset a funnel plot was produced [18, 19]. The funnel plot illustrates the number of potentially excluded studies that could have observed an effect in the opposite direction as the expected effect. The number of missing studies due to publication bias was determined using all available preclinical data. The trim-and-fill analyses were repeated using the variables that explained the largest heterogeneity to split the data into more homogeneous datasets, and determine whether the potential publication bias could be explained by the heterogeneity between studies [19]. For this analysis, only those studies with more than one level of the factor variable were used. For all trim-and-fill analyses, the R0 and L0 estimators were used to determine the number of missing studies [20]. For the p-curve analysis, the data from all the ICB treatments was used to determine whether there was evidential value of the presence of a true effect [21]. For this, the right-skewness test of the *p*-value distribution was used as described in [22].

Clinical translation

To determine whether the preclinical efficacy estimate was an accurate approximation of the corresponding clinical effect, the preclinical models were used to predict overall treatment effect and cancer-specific treatment effects. Firstly, to predict overall treatment effect, preclinical and clinical models were matched based on the treatment given. The compared estimates for both preclinical and clinical settings were derived from the models that only considered treatment as the significant modifier. A multiplicative model was used to describe the relationship between the preclinical and clinical estimates to account for the range of the ratios:

$$\widehat{y}_{Clinical} = \eta_0 * \widehat{y}^{\gamma}_{Preclinical} \tag{0.1}$$

A weighted generalized linear model with Gaussian errors and log-link function was fitted to the data using this mean model. For this, the preclinical estimates were log-transformed previously:

$$\log\left(\widehat{y}_{Clinical}\right) = \beta_0 + \beta_1 \cdot \log\left(\widehat{y}_{Preclinical}\right) \tag{0.2}$$

where $\beta_0 = \log(\eta_0)$ is the intercept, β_1 is the γ exponent in Eq. (0.1). The weights were set to be the reciprocal of the variance estimates for the preclinical effects.

This model was fit to the clinical estimates for OS and PFS separately using the two preclinical estimates derived from the HRs and MSRs, which resulted in four regression models. Only the preclinical estimates for those therapies that have a clinical implementation were used for comparison. Under the assumption of no relation between preclinical and clinical estimates, the β_1 coefficient is 0 and the corresponding *p*-value from a t-test is used to determine whether there is sufficient evidence to use preclinical estimate as predictors of clinical efficacy. A statistically significant β_1 would represent the translation constant between preclinical and clinical efficacy estimates.

To assess the cancer-specific treatment effect predictions, the preclinical model using the MSR data was matched to the clinical data based on the combination of treatment and cancer type. For this, the preclinical models using treatment and cell line as modifiers were selected. Based on the cancer cell lines' tissue of origin, they were matched with the corresponding human type of cancer using a matrix of weights with cell lines in the columns and cancer type in the rows. The weights were set so as have row sums equal to 1. The predictions from this model were compared to the PFS and OS HRs and the mean squared error (MSE) was calculated.

Furthermore, the clinical models for each efficacy measure were subjected to a leave-one-out cross-validation to calculate the prediction error. This was compared to the prediction error derived from the preclinical model to assess how well it approximated the clinical data.

Model simulations

The modifiers identified in the heterogeneity assessment were further explored to determine the magnitude of their contribution to the uncertainty of treatment effects. To determine the optimal configuration of the number of sampled cell lines and labs to choose from, a series of designs were evaluated with respect to the MSE and the power to detect significant effects. The random-effects model was simulated using the preclinical estimates of efficacy derived from the model accounting for variability due to cell line and lab effects, as well as the residual heterogeneity and the estimated cell- and lab-specific effects. A total of 60 different designs were proposed based on the combination of 1–6 different cell lines and 1–10 different studies. For each experimental design, 2000 simulations were generated in replicates of three and the estimated coefficients as well as *p*-values for the fixed effects were extracted. The estimated fixed effects from each artificial dataset were compared to the 'true' fixed effects to calculate the MSE.

Results

Internal and external validity threats are prevalent in most of the studies

Out of 226 records assessed for eligibility that met the inclusion criteria, 21 were excluded due to the sample size not being reported, other 33 were excluded since no survival was measured, and 1 was excluded due to its lack of monotherapies implemented. Other 11 publications in which the efficacy estimates could not be estimated were also excluded. 160 publications were selected based on the outlined criteria in the Methods section. These studies contain 616 experiments done in 13,811 mice, evenly split between the control (52%) and treatment groups (48%). To evaluate the internal and external validity, a subset of threats to validity in the design of experiments according to Henderson et al. [4] is presented in Table 1. Several criteria were not met in most studies: only a little over a third of the studies (37%) randomized the subjects to treatment groups, and only 5% blindly assessed the outcome to therapy (Table 1). Likewise, the determination procedure of the sample size (e.g., via power assessment) was seldom addressed, and the majority had a small group size (the median number of mice per treatment group was 10 (Table 2, Fig. 2). These flaws represent potential sources of bias in the assessment of therapeutic outcomes..

Table 1 Study characteristics of preclinical immune checkpoint blockade experiments. Numbers in parentheses represent percentages of total number of studies

Total number of studies	**160**
Total number of mice	13,811
Internal validity	
Randomization	60 (37)
Sample size justification	13 (8)
Blinded assessment of treatment effect	9 (6)
Experimental flow	8 (5)
Dose response	22 (14)
Construct validity	
Baseline characterization	5 (3)
Disease match	80 (50)
Age match	11 (7)
Mechanistic response to treatment	144 (90)
External validity	
Model replication	86 (54)
Research purpose	
Translational claim	133 (83)

Table 2 Experimental characteristics of preclinical immune checkpoint blockade treatments

Experimental design characteristics	**Median (Range)**
Mice per treatment group	10 (4–45)
Studies per treatment group	3 (1 – 79)
Cell lines per treatment group	3 (1–48)

Furthermore, there were also weaknesses in the experiments' construct validity. Most studies have a basic baseline characterization of the mouse population (i.e., either the age, sex, tumor size or stage and the inclusion/exclusion criteria are described), with only 3% of all studies characterizing all four. Furthermore, there was a mismatch of the age group of mice and the clinical populations, since most experiments (93%) used young mice (between 8–12 weeks of age). Likewise, only 22 (4%) experiments used both male and female mice, 322 (52%) used female mice, and 74 (12%) used male mice, and 198 (32%) studies did not disclose this information. In clinical trials, however, both male and female subjects are included, albeit with a higher proportion of males [23, 24]. About half of the studies explicitly matched or justified the model relevance to clinical cancers, in terms of either similar driving mutations or similar mechanisms of disease progression. Finally, most studies (90%) characterized the mechanism of response to treatment by comparison of immune cell activity in treatment versus control groups.

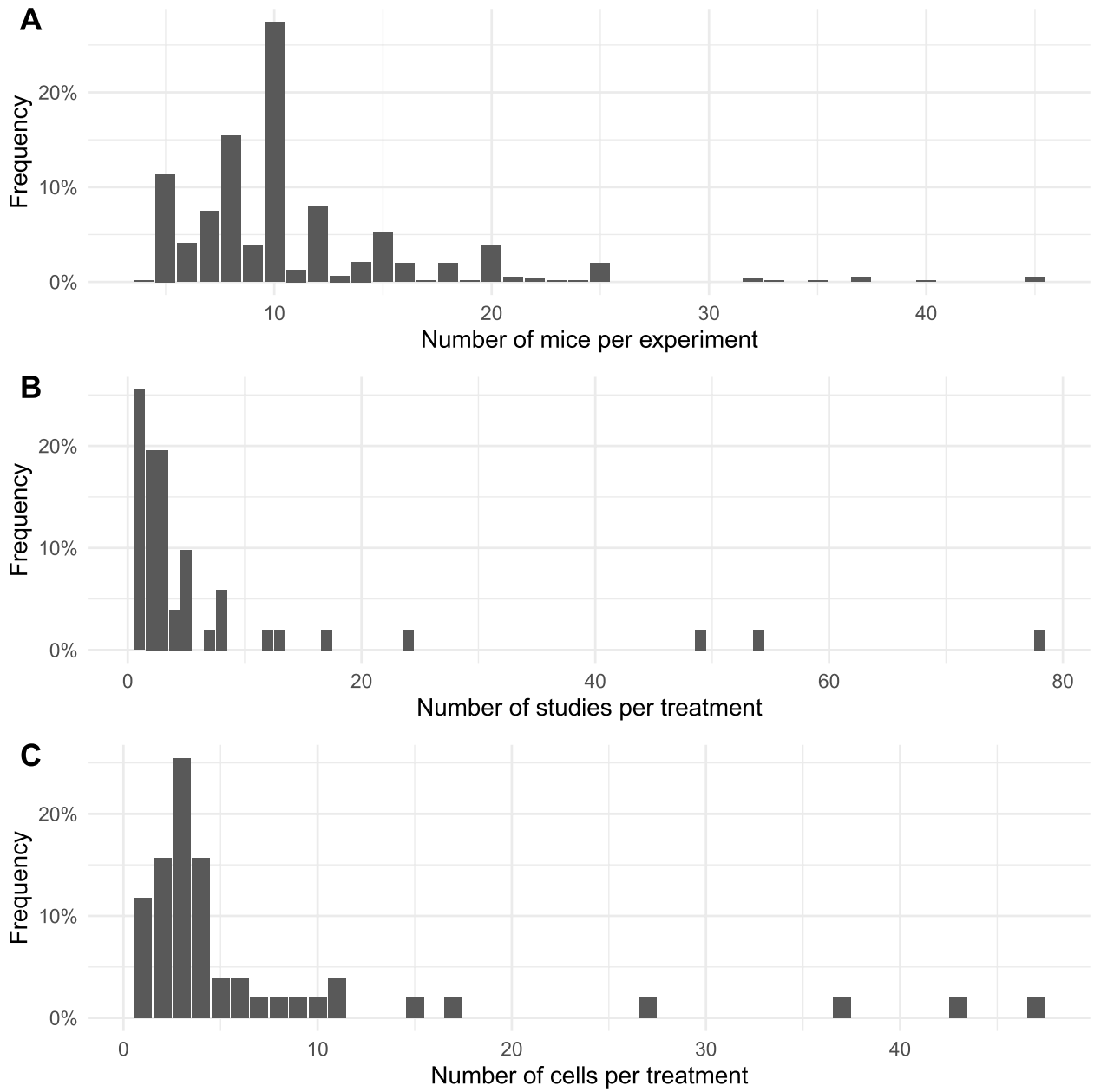

Fig. 2 Experimental design characteristics. The figure shows the distribution of the design variables presented in Table 2

The external validity of these preclinical experiments was limited in scope, since only half of the studies used more than one tumor model to investigate the response to treatment. Furthermore, there were no independent replicates by other research groups mentioned in any of them. Regardless of these flaws, 83% of the publications claim some translational property with regards to the clinical application of the investigated drug or therapy.

Trim-and-fill analysis shows missing preclinical studies for the hazard ratio but not for the median survival ratio

To determine whether the internal or external validity threats influenced the overall preclinical estimates of efficacy, a mixed-effects model was fitted with treatment and each validity threat variable as factors. However, neither the randomization, blinded assessment of outcome nor the model replication had any significant effect on the estimates (HR *p*-values: 0.84, 0.90, 0.67; MSR *p*-values: 0.94, 0.2125, 0.8025, respectively).

Next, the potential presence of publication bias was assessed using two methods: the trim-and-fill and the p-curve analyses. Both methods were employed on the two measures of efficacy, namely the HRs and MSRs, given their differences in distribution (Supplementary Fig. 2). The MSR had more efficacy estimates closer to the nonsignificant value of 1 in comparison to the HR, which had many estimates well below 1 and thus declared significant. This overdispersion could be due to the violation of the model assumptions of the HR calculated with the Cox proportional hazards model and the small number of mice used in the experiments (Table 2). For the trim-and-fill method, separate analyses were performed for each treatment due to the influence that inter-study heterogeneity can exert on the results. The results of this method are shown in Figs. 3 and 4. For the case of HRs, most treatments showed a significant number of missing studies, especially in the monotherapy groups (Fig. 3). This was not the case for MSRs, where apparently there is no publication bias in the clinically positive direction (Fig. 5). This discrepancy could be explained by the difference in the distribution in the data, as the HRs had a longer right tail and values clustered below the non-significant value of 1 (Supplementary Fig. 2). On average, the treatment effect was overestimated by 12% in the HRs due to publication bias.

As previously stated, the trim-and-fill method can falsely output a number of missing studies when there is considerable inter-study heterogeneity [25]. For this reason, the p-curve analysis was employed, which uses the distribution of *p*-values derived from the hypotheses tests of the effect sizes [21]. This distribution is subjected to tests of right-skewness and flatness to determine whether the data contains evidential value, i.e., that the observed effect is not spurious due to p-hacking or data-mining [26]. This analysis was performed on the subset of treatments of interest, namely anti-CTLA-4, anti-PD-1, anti-PD-L1, antiCTLA-4 and their two-way combinations with each other and with chemotherapy. The curves for both efficacy measures, as well as the results of the right-skewness and flatness tests are shown in Fig. 5.

The p-curves for both measures indicate that there is evidential value in the preclinical data for the efficacy of ICB (test for right-skewness is significant). Thus, it is unlikely that there was any p-hacking or data mining to produce results that are statistically significant. Combined with the previous analysis' results, this indicates that, although there were some potential missing studies identified in the reporting of HRs, the reported studies that were statistically significant provided sufficient evidence of true efficacy. In the following section, the potential influence of inter-study heterogeneity on the potential missing studies will be explored.

The effects of the study and cell line explain the most variation in preclinical treatment effects

To identify potential sources of variability between studies of the same treatment, several experimental design variables were considered based on published knowledge of their impact on treatment effect. These included the cell line [13], sex [27] treatment start day [28, 29], and site of tumor injection [30]. Other design variables of interest included the dose and number of initial tumor cells injected. Additionally, the internal validity variables qualifying the randomization and blinded assessment were considered too. Finally, the variables for the institute, lab, and publication study in which the experiments were performed and reported were also included (Table 3).

The model with only treatments as modifiers explained 48% of the variability in MSRs and 15% in HRs. Only the study combined with the treatments could completely explain the heterogeneity between experiments in the MSRs; in the HRs, these variables explained 46% of the observed heterogeneity. Although none of the random effects in combination with the treatments completely explained the heterogeneity in HRs, the cell line, institute, and lab explained the greatest percentages of the variability in HRs (Table 3). Although the dose itself was not a significant factor for explaining inter-study heterogeneity, the possible existence of a dose–response relationship was further examined for the monotherapies of anti-CTLA-4, anti-PD-1, and anti-PD-L1 (Supplementary Figs. 3 and 4). However, there was no significant association between the dose (mg/kg) and the effect size in either measure of efficacy for any therapy.

Furthermore, there was considerable residual heterogeneity in the model fitted to the HRs, which could not be further reduced by including the study and cell line as modifiers in the same model (HRs: p-value: < 0.0001). This heterogeneity could potentially have an effect on the previous assessment of publication bias [19]. To determine whether the residual heterogeneity could explain the publication bias previously identified, the dataset was partitioned into subsets according to the variables that accounted for the largest percentages of variability between studies, namely the cell and the study. Only subsets with more than two observations were kept, and therefore some cell lines or studies with only one estimate were not included in this analysis. The mixed effects models were fit to the HRs in each of these subsets and the trim-fill method was implemented to determine whether in these more homogeneous datasets there were still potentially missing studies. Although the estimates changed between the different subsets for all treatments, there were still a significant and equivalent number of missing studies in each subset compared to the whole dataset (Supplementary Figs. 5 – 10). Thus, it is unlikely

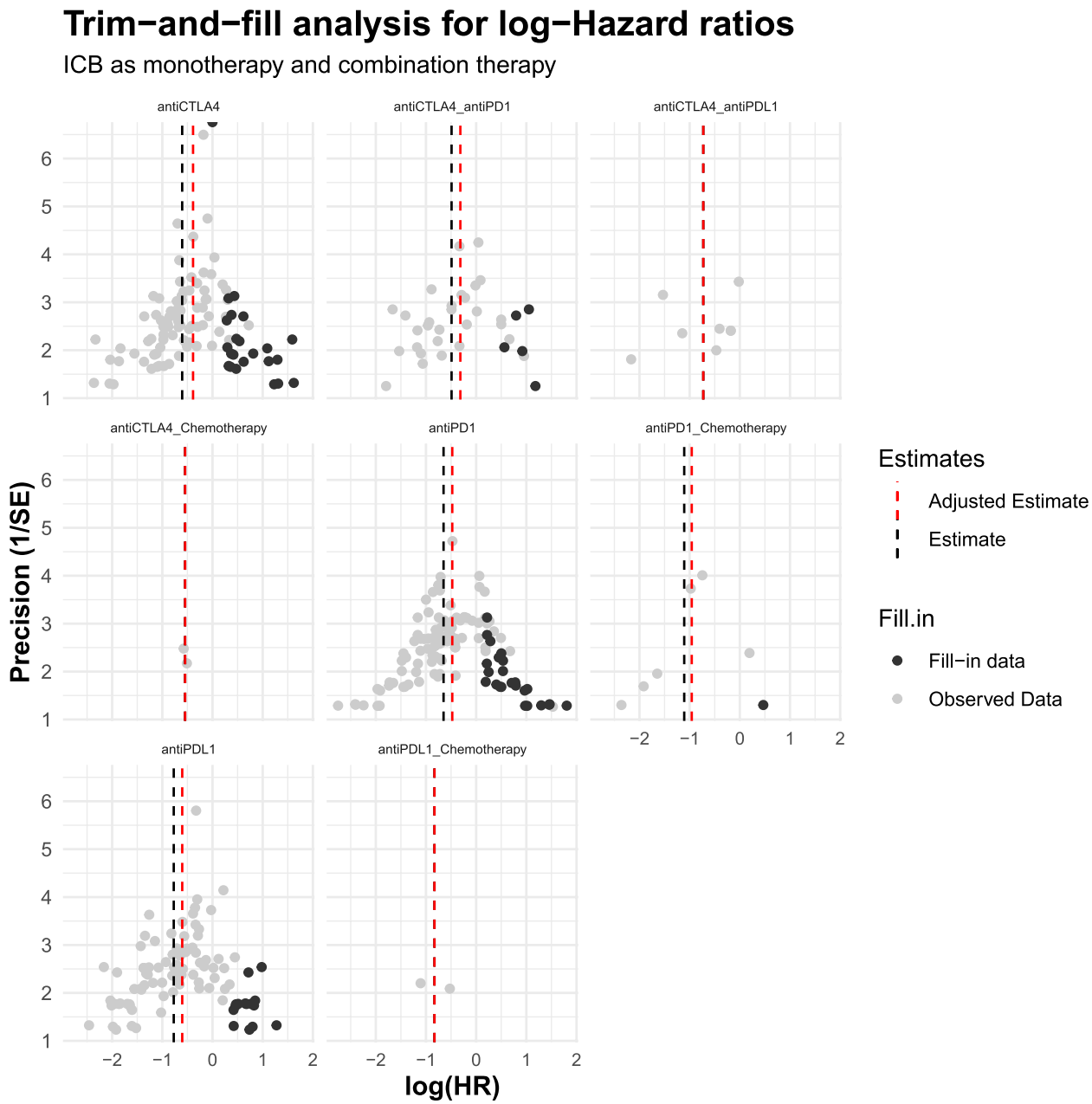

Fig. 3 Funnel plots of log-HRs in preclinical experiments of ICB. For each treatment, a trim-and-fill analysis was conducted to determine the number of missing studies due to potential publication bias. The black dots represent missing studies, and the red dotted line the adjusted estimate taking those missing experiments into consideration. Figure shows a subset of treatments from the total 53 of preclinical therapies

that the potentially omitted studies were a product of inter-study heterogeneity.

Furthermore, most of the studies used only one tumor cell line to quantify treatment effect (Table 2) and therefore the obtained estimate is particularly susceptible to variation stemming from either the cell or study variables. To understand how the experimental choices of these two factors contributed to the accuracy and precision of the HR and MSR estimates, a simulation experiment was performed. With this purpose, the model estimates of the treatment effects as well as the cell and study effects were used to randomly generate artificial datasets with different combinations of cell lines (1–6 different cell lines) and studies (1–10 different studies). This constituted 60 different experimental designs, simulated in replicates of 3, with 2000 datasets for each design. Overall, 360,000

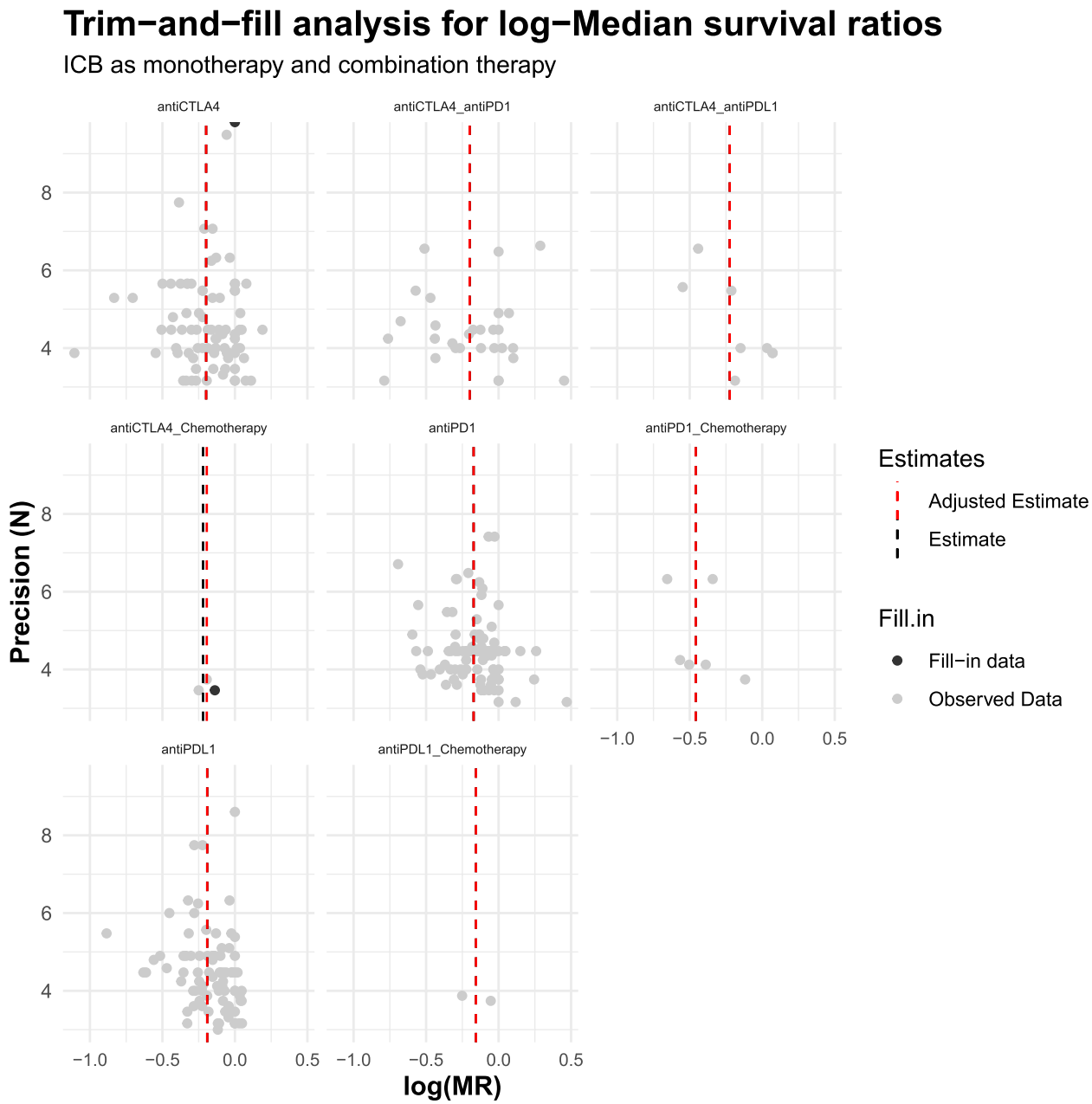

Fig. 4 Funnel plots of log-MSRs in preclinical experiments of ICB. For each treatment, a trim-and-fill analysis was conducted to determine the number of missing studies due to potential publication bias. The black dots represent missing studies, which in this setting only two were found. Figure shows a subset of treatments from the total 53 of preclinical therapies

simulations were run and from each simulation, the treatment effects and the *p*-value were extracted. From these, the mean squared error (MSE) with respect to the overall estimated treatment effects (Fig. 6) and the power to detect a significant effect (Supplementary Figs. 11 and 12) were calculated for each treatment.

In the HRs, the MSE showed significant variation between treatments consistently over all experimental designs, with the combination for anti-CTLA-4 and anti-PD-L1 having the largest MSE. The combination of chemotherapy with either anti-PD-1 and anti-PD-L1 showed the smallest MSE among all treatments. Interestingly, for all treatments, the MSE only stabilized when using 6 cell lines and 3 studies. In contrast, the MSE from MSRs were much smaller compared to the HRs' MSE and, for all treatments, there was a steep descent in MSE when using 3 studies and anywhere from 2–4 cell lines.

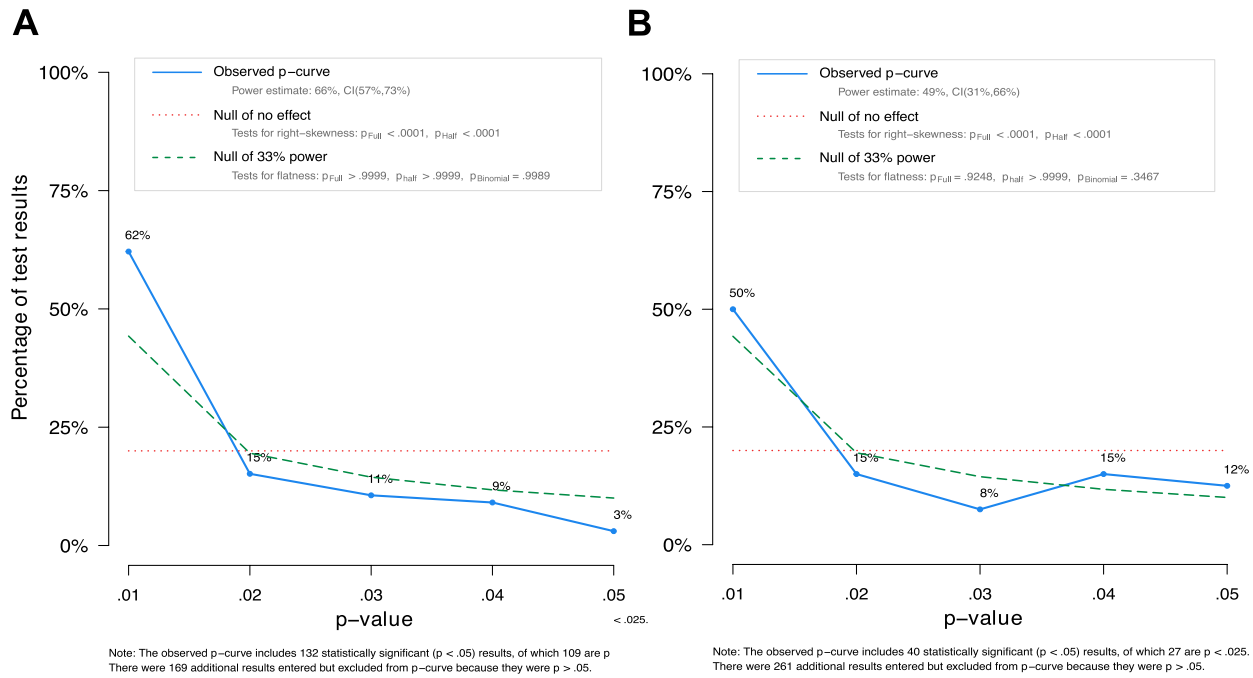

Fig. 5 p-curve analysis plots of log-HRs (**A**) and log-MSRs (**B**) in preclinical experiments of ICB. For each treatment, a p-curve analysis was conducted to determine the number of missing studies due to potential publication bias. Figure shows a subset of treatments from the total 53 of preclinical therapies: anti-CTLA-4, anti-PD-1, anti-PD-L1, anti-CTLA-4 + anti-PD-1, anti-CTLA-4 + anti-PD-L1, anti-CTLA-4 + Chemotherapy, anti-PD-1 + Chemotherapy, and anti-PD-L1 + Chemotherapy

For all treatments, the statistical power lay above 0.8 for all combinations of the random effects in both HRs and MSRs. Thus, the probability to detect a statistically significant effect was sufficient even for designs with one cell line and one study only (Supplementary Figs. 11 and 12).

Table 3 Assessment of heterogeneity in efficacy measures of preclinical ICB experiments

Modifier	Log-HRs		Log-MSRs	
	AIC	R^2	AIC	R^2
Study	991	46	329	99
Lab	1004	39	274	78
Institute	1016	34	235	65
Strain	1050	17	31	64
Cell	1025	28	224	57
Number of tumor cells injected	1047	17	25	52
Day of treatment start after tumor inoculation	1052	15	27	51
Cancer type	1038	20	86	50
Sex	1043	15	37	48
Randomization	1055	14	31	48
Blinded assessment	1055	15	32	48
Site of injection	1057	13	47	47
Route of administration	1036	22	59	47
Dose	1050	15	33	47

Clinical efficacy estimates show heterogeneity, partially explained by the cancer type

Clinical phase III studies of ICB as monotherapy or combination therapy were collected based on the strategy outlined in the Methods section. Like in the preclinical setting, a univariate random effects linear model was fitted to the clinical data to analyze the potential publication bias and heterogeneity between studies. In the OS HRs, there was a potential publication bias present as well, especially in the anti-PD-1 therapy (Supplementary Fig. 13). Although the rest of the treatments did not exhibit a potentially large publication bias, the small number of points used for the analyses preclude any meaningful assessment. However, in the PFS HRs, there was no significant potential omission of any studies (Supplementary Fig. 14). Furthermore, the p-curve analysis also provided evidential value for the presence of a true effect in both efficacy measures (Supplementary Fig. 15).

The treatment-specific effects in clinical studies were estimated similarly to the preclinical setting, testing various modifiers to determine potential sources of heterogeneity between studies. Although the cancer type could explain some of the heterogeneity in both OS and PFS HRs ($R^2 = 46$ and 36%, respectively), there was still residual variability between studies (p-value = 0.0008, < 0.0001)

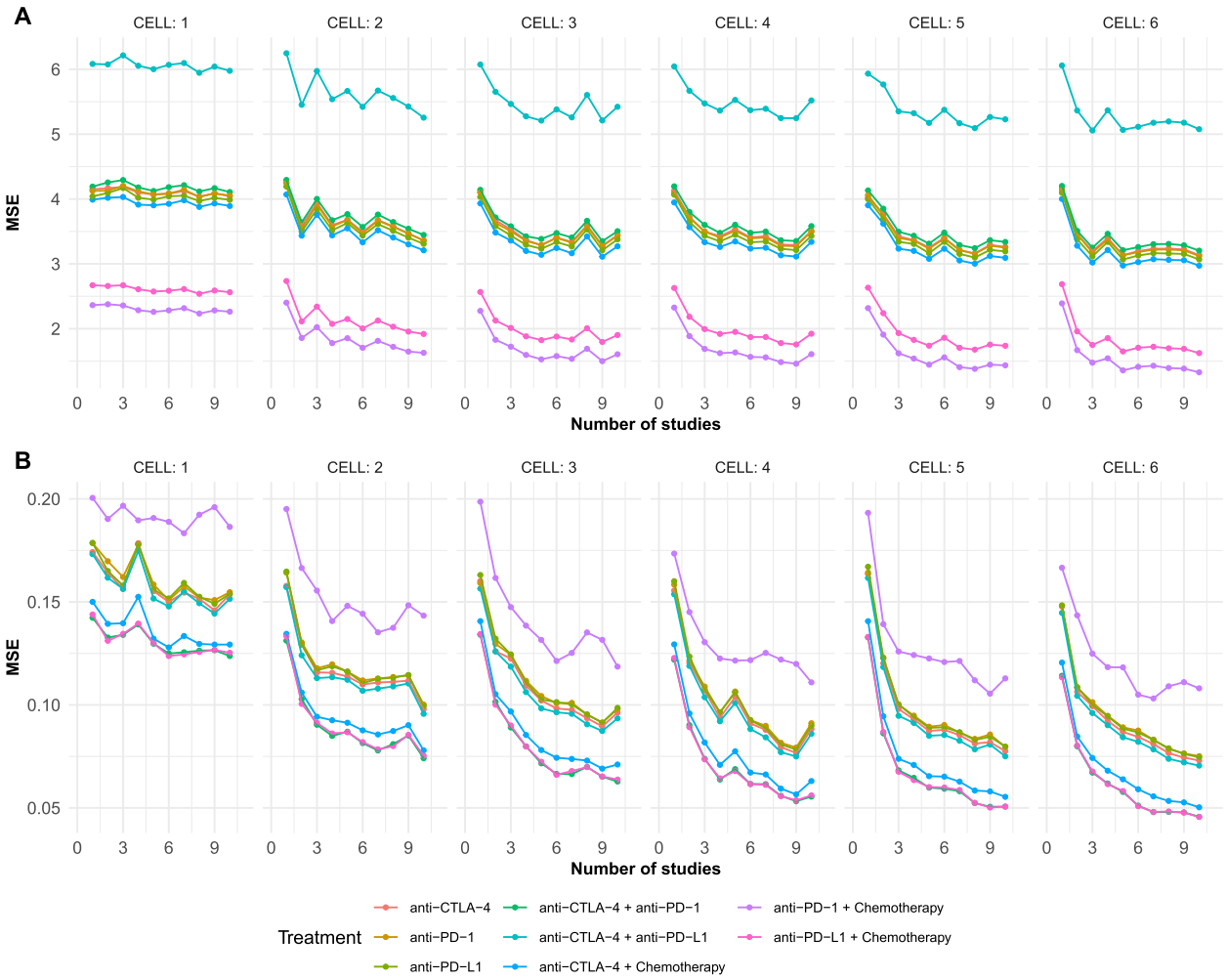

Fig. 6 Mean squared error from experimental designs of ICB treatments. Results from the simulation experiments of 60 experimental designs with 500 simulations per design and 3 replicates per simulation. Shown are the resulting mean squared errors with respect to the estimated treatment effects for either (**A**) the HRs or the (**B**) MSRs. Each panel represents the number of cell lines used

(Table 4). In both measures, the variability could be reduced by additionally including the type of masking as a modifying variable (for OS: $R^2 = 70.25\%$, p-value $= 0.0415$) (for PFS: $R^2 = 73.53\%$, p-value $= < 0.0001$), but there was still residual heterogeneity.

The preclinical model predictions of clinical efficacy approximated the observed estimates in a cancer and treatment-specific manner

The overall preclinical treatment effect estimates for both efficacy measures of survival were compared to the overall clinical treatment effects estimates of the OS and PFS HRs. For this, the treatment effects were estimated from the model using only the treatments as the relevant modifiers for each efficacy measure in both preclinical and clinical settings. Comparing the preclinical to the clinical fixed effects, there was a clear overestimation of the latter by the former for HRs: The preclinical HRs overestimated the clinical OS and PFS HRs by 65% on average. In MSRs, there was a slight overestimation of effect in a few of the treatment estimates, but on average, this overshoot was mild in comparison (8–12%). Furthermore, for the fixed

Table 4 Heterogeneity assessment in clinical studies of immune checkpoint inhibitors. The Akaike Information Criterion (AIC) and the percentage of heterogeneity explained by each modifier (R^2) are shown for each model

Modifier	Log-OS HRs		Log-PFS HRs	
	AIC	R^2	AIC	R^2
Cancer type	-8.134	46%	51.85	37%
Therapeutic agent(s)	22.17	0%	65.08	12%
Type of masking	-30.38	27%	45.8	46%

effects of interest, there was a lack of association between preclinical and the clinical estimates since both measures of preclinical efficacy failed to predict the observed drug effects in OS (Supplementary Figs. 16 and 17) and PFS (Fig. 7 and Supplementary Fig. 18). The t-test for the relationship between the preclinical and clinical estimates for all comparisons, i.e., that γ was not significantly different than 0, did not generate any significant results: preclinical HR vs. clinical OS HR, p-value = 0.83; preclinical HR vs clinical PFS HR: p-value = 0.66; preclinical MR vs clinical OS HR, p-value = 0.08; preclinical MR vs clinical PFS HR: p-value = 0.58.

The preclinical estimates from MSRs showed large uncertainty in some treatments (anti-CTLA-4+anti-PD-L1, anti-CTLA-4+Chemotherapy, and anti-PD-L1+Chemotherapy), partly due to the small number of studies for those therapies (Table 5). However, for the treatments with higher precision, such as in anti-CTLA-4, anti-PD-1, anti-PD-L1, and anti-CTLA-4+anti-PD-1, the preclinical estimates were quite close to the clinical OS HRs estimates. In the case of preclinical HRs, they were consistently more favorable than the clinical estimates. Additionally, in most treatments the prediction intervals derived from preclinical data were completely out of range of the estimated clinical confidence intervals; only in the combination of anti-CTLA-4 with anti-PD-1, and the combinations of anti-CTLA-4 and anti-PD-1 with chemotherapy there was some agreement. However, this overlap seemed to be due to the large uncertainty in preclinical estimates rather than a partial agreement with the clinical efficacy estimate.

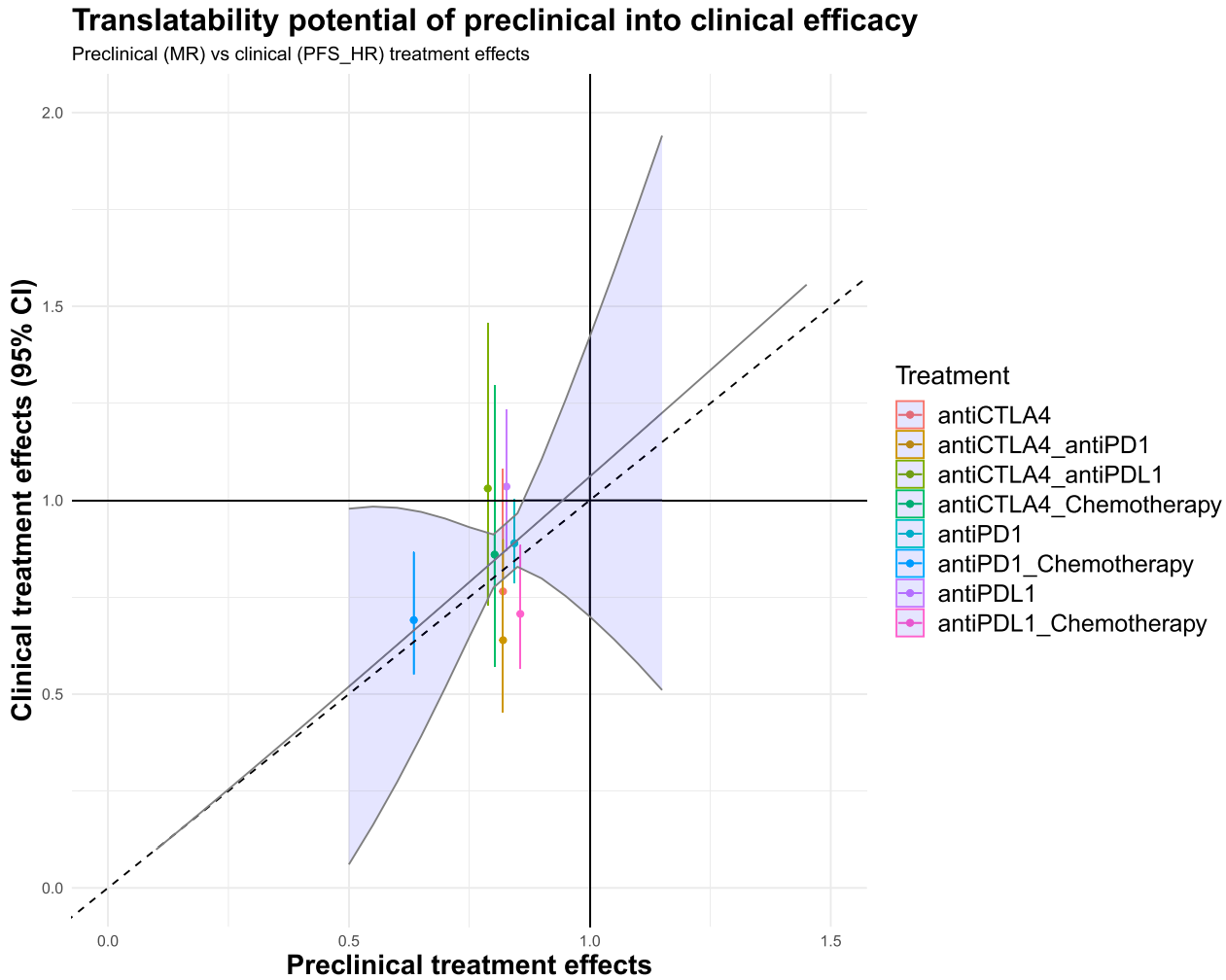

Fig. 7 Comparison of preclinical vs. clinical PFS fixed effects of ICB therapies. The preclinical (MSR) and clinical (PFS HR) estimates from the meta-analyses for each treatment were plotted against each other (dots) along with their 95% confidence intervals. The red dotted line represents the slope of the regression passing through the origin of the clinical efficacy estimates on the preclinical ones and its 95% confidence interval. The black dotted line represents the perfect correspondence between preclinical and clinical estimates of efficacy. This figure corresponds to the data presented in Table 5

Table 5 Preclinical (MSR and HR) and clinical (OS and PFS HR) estimates of survival efficacy. LB: lower bound; UB: upper bound

Treatment	Preclinical MSRs			Preclinical HRs			Clinical OS HRs			Clinical PFS HRs		
	Estimate	95% CI		Estimate	95% CI		Estimate	95% CI		Estimate	95% CI	
		LB	UB		LB	UB		LB	UB		LB	LB
Anti-CTLA-4	0.82	0.78	0.86	0.30	0.17	0.53	0.88	0.76	1.01	0.76	0.54	1.08
Anti-PD-1	0.84	0.80	0.88	0.31	0.18	0.55	0.77	0.73	0.82	0.89	0.79	1.00
Anti-PD-L1	0.83	0.79	0.87	0.30	0.17	0.54	0.86	0.79	0.93	1.04	0.87	1.23
Anti-CTLA-4 + anti-PD-1	0.82	0.75	0.89	0.36	0.20	0.66	0.71	0.62	0.81	0.64	0.45	0.90
Anti-CTLA-4 + anti-PD-L1	0.79	0.67	0.93	0.36	0.20	0.66	0.87	0.74	1.03	1.03	0.73	1.46
Anti-CTLA4 + Chemotherapy	0.80	0.54	1.20	0.31	0.10	0.99	0.87	0.75	1.01	0.86	0.57	1.30
Anti-PD-1 + Chemotherapy	0.63	0.53	0.76	0.17	0.08	0.38	0.73	0.66	0.82	0.69	0.55	0.87
Anti-PD-L1 + Chemotherapy	0.86	0.58	1.25	0.20	0.07	0.60	0.82	0.73	0.91	0.71	0.56	0.88

The next question was whether cancer-type specific predictions could improve the prediction of the observed clinical effects. For this, the preclinical models for both efficacy measures with the treatments and cell lines as modifiers were used to predict efficacy in various clinical trials of ICB. As described in the Methods section, the preclinical models were used to generate cancer type specific predictions of clinical efficacy (see Supplementary Table 1). These were matched to the clinical OS and PFS HRs observed in multiple trials of ICB and the prediction MSE was compared to the leave-one-out cross-validation (LOO CV) MSE derived from the clinical model. For both types of clinical efficacy, the prediction MSE was close to the cv MSE when using the model trained with the MSR (Table 6 and Supplementary Figs. 19 and 20). There was a large bias in the predictions derived from preclinical HRs, in which all consistently produced more favorable results than the ones observed in the clinic (Supplementary Figs. 21 and 22).

Discussion

The relevance of preclinical results for translational purposes in clinical efficacy has been a topic of debate due to the concerns regarding the internal and external validity of the experiments. In general, the lack of bias control strategies in experimental design generates efficacy estimates with a larger magnitude in the clinically positive direction than would be expected when controlling for bias. The studies of ICB examined here exhibited design flaws that made them vulnerable to internal and external validity threats in the prevention of bias.

Table 6 Prediction MSE from the clinical and preclinical models. The leave-one-out cross-validation error is shown for the clinical model. For the preclinical models, the errors derived from the cancer-type specific predictions of the corresponding measures of clinical efficacy are shown for each of the preclinical variables considered (MSR and HR)

Model	OS HR MSE	PFS HR MSE
Clinical CV	0.0184	0.0836
Preclinical MSR	0.0221	0.0901
Preclinical HR	0.1181	0.2595

Internal validity

The internal validity of many studies was threatened by the lack of randomization, blinded assessment, sample size determination and optimal dose finding. This is not limited to preclinical studies in ICB, but is a larger issue across multiple therapeutic fields [6–9]. Furthermore, sample size and dose finding were seldom addressed, and most studies used a small number of mice with a predetermined dosing regimen. However, the inter-individual variation in mice was shown to be a relevant factor in the response to treatment for different immunotherapies, even for mice with similar genetic backgrounds and housing conditions [31, 32]. Thus, finding a sample size that contemplates this source of variation is key in minimizing the estimates' bias. It is likely that the preclinical bias in efficacy estimates contributed to the difference between the individual preclinical and clinical efficacy estimates. Nevertheless, other sources for this discrepancy also include the differences between species and the threats to construct validity, i.e., the choice of control groups in animal vs human studies. Further studies with unbiased estimates of preclinical studies could better assess the contribution of each of these factors to the overestimation of clinical efficacy.

External validity

In preclinical experiments of ICB, the external validity was threatened by the lack of multiple tumor models and animal species tested, as well as flaws in the construct validity (i.e., matching of age, disease mechanism, and baseline characterization to clinical patients). Specifically,

there was large variation in efficacy measures due to the cell line and study in which the experiments were performed. The simulations with different experimental designs allowed for the identification of optimal combinations of the number of cell lines and studies to obtain reliable preclinical efficacy estimates. For MSRs, a combination of 2 cell lines and 3 studies was adequate, however, for HRs, a combination of 6 cell lines and 3 studies was required for MSE minimization. A different study found similar effects of the lab on the effect size estimates, and via simulation came to a recommendation of 2–4 labs to fully account for the inter-lab variability [12]. Although not explored in this work, the within-study sampling error is another potential source of significant variation between efficacy estimates. A way to tackle this is through multi-batch experiments, which minimize the contribution of unaccounted sources of variability to the effect size estimate [33].

Another source of bias was the outcome measure chosen to compare the treatment groups. In this meta-analysis, all studies used the logrank test to determine whether the treatment effect is statistically significant in improving survival with respect to the control group. In the case of two-way comparisons, the logrank test is equivalent to the test for the HR from the Cox proportional hazards model [34]. This model quantifies the relative hazard of observing an event (e.g., death) between two different groups. However, some of the key assumptions of the model, such as non-zero survival curves and proportionality between baseline hazard and the treatment-related hazard, were violated in preclinical experiments. This led to HR estimates with large bias, which, compared to the log MSR, were worse at approximating the human clinical efficacy. The MSR on the other hand had a smaller bias and was better at approximating the observed clinical efficacy estimates.

Clinical translation

There are differences between species that some have claimed to be insurmountable for the purpose of extrapolating the results to the clinic [5]. Particularly in cancer immunotherapy there are differences in immune system functions [35], in the monoclonal antibodies used in humans versus mice [36], in disease progression [13], among others. Assessing how significant these differences are towards predicting clinical efficacy is non-trivial, not only because of the complexity of the immune response but also due to the variation between individuals and tumor models [13, 37, 38]. Indeed, when comparing the group of overall preclinical efficacy estimates to their clinical counterparts, there was almost no association in most comparisons. In fact, there was only a weak association between the MSR and PFS HRs, which means that the overall preclinical estimates of efficacy could not be used to predict overall clinical outcomes for the treatments considered here. However, it should be noted that a small sample size was available due to the reduced number of therapies currently implemented in the clinic. Thus, it cannot be concluded that there is no association between preclinical and clinical efficacy estimates for any possible treatment, only for those here examined. As more combinations of ICB are tested in the clinic, this model could be revisited in the future to give a better assessment. On the other hand, the cancer-specific predictions of treatment efficacy were better approximations of the observed clinical effect. The prediction error was close to the cross-validation error from the clinical model, thus making the preclinical model trained with MSR data useful for predicting clinical efficacy in ICB trials given the current knowledge. However, it should be noted that the clinical model had significant residual heterogeneity that could not be explained by any of the considered variables. Indeed, other studies have identified other factors as relevant in the efficacy, e.g., the mutational landscape of the tumor, PD-L1 status, and intratumor immune cell composition [39–42]. In this study, the efficacy data was from the intention-to-treat group in each study, hence these factors could vary significantly between study groups and potentially explain the residual heterogeneity of the model. Although the residual heterogeneity might put into question the validity of the clinical model for both measures, it is still useful in describing the effect of the cancer type and treatment in the efficacy estimate. Further work should focus on incorporating those relevant factors affecting clinical efficacy to improve both the preclinical and clinical models.

Study limitations

This is not a systematic review of the up-to-date literature in preclinical and clinical studies, and a significant number of more recent preclinical experiments was left out for this meta-analysis. Another important limitation of this study is that it mostly focused on SyMM, thus it remains unknown how adequate other mouse models such as GEMM and humanized mice are at approximating human clinical efficacy in ICB. Furthermore, although dose was used in the examination of possible sources of heterogeneity between preclinical studies, there was considerable variability in the dose frequency and duration of treatment between studies. Therefore, the effect of dose scheduling needs to be further evaluated. Similarly, in the clinical meta-analysis only the drug type was considered to summarize treatment effect, but there was variation in doses, especially for ipilimumab (anti-CTLA-4 drug: 3 mg/kg – 10 mg/kg) which has shown to influence its efficacy [43]. In both the preclinical and clinical

meta-analyses, the different mAbs types were combined to estimate the treatment effect based on the drug type, i.e., anti-PD-1, anti-PD-L1 and anti-CTLA-4. This was implemented to compare between preclinical and clinical estimates more easily. In the clinic, other meta-analysis have found no significant differences in OS or PFS for PD-1 and PD-L1 mAbs [44, 45]. However, the mouse mAbs for the same target have shown variation of effect, partly due to the IgG type [46, 47]. Finally, different mAbs were used in the mouse experiments and the clinic due to the lack of cross-reactivity, which have shown additional effects to receptor blockade (i.e. effector cell depletion) affecting their efficacy [36].

Conclusions

Overall, the preclinical experiments had several flaws in the minimization of threats to internal and external validity. These had marked effects on the estimates of preclinical efficacy by making them more favorable than would be expected in a better experimental design. Furthermore, the most used preclinical efficacy measure (HR) also skewed the efficacy towards a more clinically positive direction. A different measure (MSR) should be used to avoid such bias, as it was shown to be more predictive of clinical efficacy when considering various cell types that can be matched to their respective human cancer type.

Abbreviations

AIC	Akaike information criterion
GEMM	Genetically engineered mouse model
HR	Hazard ratio
ICB	Immune checkpoint blockade
MSE	Mean squared error
MSR	Median survival ratio
PFS	Progression-free survival
SyMM	Syngeneic mouse model

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s41231-023-00151-x.

Additional file 1: Supplementary Figure 1. Comparison of reported vs. calculated median survival. **Supplementary Figure 2.** Differences in distribution between HRs and MSRs. **Supplementary Figure 3.** Dose-response relationship in HRs. **Supplementary Figure 4.** Dose-response relationship in MSRs. **Supplementary Figure 5.** Trim-fill analysis by study of log-HR efficacy estimates in anti-CTLA-4-treated mice. **Supplementary Figure 6.** Trim-fill analysis by study of log-HR efficacy estimates in anti-PD-1-treated mice. **Supplementary Figure 7.** Trim-fill analysis by study of log-HR efficacy estimates in anti-PD-L1-treated mice. **Supplementary Figure 8.** Trim-fill analysis by cell line of log-HR efficacy estimates in anti-CTLA-4 – treated mice. **Supplementary Figure 9.** Trim-fill analysis by cell line of log-HR efficacy estimates in anti-PD-1-treated mice. **Supplementary Figure 10.** Trim-fill analysis by cell line of log-HR efficacy estimates in anti-PD-L1- treated mice. **Supplementary Figure 11.** Statistical power from experimental designs of ICB therapies for HRs. **Supplementary Figure 12.** Statistical power from experimental designs of ICB therapies for MSRs. **Supplementary Figure 13.** Trim-and-fill analysis by treatment of clinical OS HRs. **Supplementary Figure 14.** Trim-and-fill analysis by treatment of clinical PFS HRs. **Supplementary Figure 15.** P-curve analysis for clinical measures of efficacy. **Supplementary Figure 16.** Comparison of preclinical vs. clinical OS fixed effects of ICB therapies. **Supplementary Figure 17.** The preclinical (HR) and clinical (OS HR) estimates from the meta-analyses for each treatment were plotted against each other (dots) along with their 95% confidence intervals. **Supplementary Figure 18.** Comparison of preclinical vs. clinical PFS fixed effects of ICB therapies. **Supplementary Figure 19.** Preclinical model predictions of clinical efficacy in PFS HRs. **Supplementary Figure 20.** Preclinical model predictions of clinical efficacy in OS HRs. **Supplementary Figure 21.** Preclinical model predictions derived from HRs vs. clinical PFS HRs from various clinical studies of ICB. **Supplementary Figure 22.** Preclinical model predictions derived from HRs vs. clinical OS HRs from various clinical studies of ICB. **Supplementary Table 1.** Preclinical model cancer-type specific predictions of treatment efficacy derived from MSR. **Supplementary Figure 23.** Leave-one-out cross validation predictions of median survival.

Acknowledgements

The authors would like to thank Bárbara Rocha Loza for helping in the screening of publications and collection of data. Also, thanks to Christoph Niederalt for his feedback in preparing this manuscript.

Authors' contributions

JMTP, JL and RB were responsible for the study conception and design; JMTP retrieved and analyzed the data, and created the code; JMTP, JL, RB and CS interpreted the study results; JMTP redacted an initial draft and JL, RB and CS revised it.

Funding

This research was funded by the European Union Marie-Curie Innovative Training Network ENLIGHT-TEN under grant agreement 675395.

Availability of data and materials

The datasets supporting the conclusions of this article are included as well as the code is available in the GitHub repository: Survival Meta-Analysis.

Declarations

Ethics approval and consent to participate

The data gathered and analyzed in this study stemmed entirely from published preclinical and clinical studies, thus, no ethics approval was necessary.

Consent for publication

Not applicable.

Competing interests

JMTP is a former Bayer employee. CS, RB, and JL are employees at Bayer.

Received: 10 April 2023 Accepted: 4 July 2023
Published online: 15 August 2023

References

1. Kather JN, Berghoff AS, Ferber D, Suarez-Carmona M, Reyes-Aldasoro CC, Valous NA, et al. Large-scale database mining reveals hidden trends and future directions for cancer immunotherapy. OncoImmunology. 2018;7(7):e144412.
2. Denayer T, Stöhr T, Van Roy M. Animal models in translational medicine: Validation and prediction. New Horiz Transl Med. 2014;2(1):5–11.
3. Li QX, Feuer G, Ouyang X, An X. Experimental animal modeling for immuno-oncology. Pharmacol Ther. 2017;173:34–46.
4. 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(7):e1001489.

5. Pound P, Ritskes-Hoitinga M. Is it possible to overcome issues of external validity in preclinical animal research? Why most animal models are bound to fail. J Transl Med. 2018;16(1):304.
6. van der Worp HB, Howells DW, Sena ES, Porritt MJ, Rewell S, O'Collins V, et al. Can Animal Models of Disease Reliably Inform Human Studies? PLOS Med. 2010;7(3):e1000245.
7. 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. Teare MD, editor. eLife. 2015;4:e08351.
8. Watzlawick R, Antonic A, Sena ES, Kopp MA, Rind J, Dirnagl U, et al. Outcome heterogeneity and bias in acute experimental spinal cord injury: A meta-analysis. Neurology. 2019;93(1):e40-51.
9. Mattina J, MacKinnon N, Henderson VC, Fergusson D, Kimmelman J. Design and Reporting of Targeted Anticancer Preclinical Studies: A Meta-Analysis of Animal Studies Investigating Sorafenib Antitumor Efficacy. Cancer Res. 2016;76(16):4627.
10. Sena ES, van der Worp HB, Bath PMW, Howells DW, Macleod MR. Publication Bias in Reports of Animal Stroke Studies Leads to Major Overstatement of Efficacy. PLOS Biol. 2010;8(3): e1000344.
11. Laajala TD, Jumppanen M, Huhtaniemi R, Fey V, Kaur A, Knuuttila M, et al. Optimized design and analysis of preclinical intervention studies in vivo. Sci Rep. 2016;02(6):30723.
12. Voelkl B, Vogt L, Sena ES, Würbel H. Reproducibility of preclinical animal research improves with heterogeneity of study samples. PLOS Biol. 2018;16(2):e2003693.
13. Mosely SIS, Prime JE, Sainson RCA, Koopmann JO, Wang DYQ, Greenawalt DM, et al. Rational Selection of Syngeneic Preclinical Tumor Models for Immunotherapeutic Drug Discovery. Cancer Immunol Res. 2017;5(1):29.
14. Vesterinen HM, Sena ES, Egan KJ, Hirst TC, Churolov L, Currie GL, et al. Meta-analysis of data from animal studies: A practical guide. J Neurosci Methods. 2014;15(221):92–102.
15. Berkey CS, Hoaglin DC, Mosteller F, Colditz GA. A random-effects regression model for meta-analysis. Stat Med. 1995;14(4):395–411.
16. Viechtbauer W. Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. J Educ Behav Stat - J EDUC BEHAV STAT. 2005;1(30):261–93.
17. Hedges L, Olkin I. Statistical Methods in Meta-Analysis. Stat Med. 1985.
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.
19. Sterne JAC, Sutton AJ, Ioannidis JPA, Terrin N, Jones DR, Lau J, et al. Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. BMJ. 2011;22(343):d4002.
20. Shi L, Lin L. The trim-and-fill method for publication bias: practical guidelines and recommendations based on a large database of meta-analyses. Medicine (Baltimore). 2019;98(23):e15987.
21. Simonsohn U, Nelson LD, Simmons JP. p-Curve and Effect Size: Correcting for Publication Bias Using Only Significant Results. Perspect Psychol Sci J Assoc Psychol Sci. 2014;9(6):666–81.
22. Simonsohn U, Simmons JP, Nelson LD. Better P-curves: Making P-curve analysis more robust to errors, fraud, and ambitious P-hacking, a Reply to Ulrich and Miller (2015). J Exp Psychol Gen. 2015;144(6):1146–52.
23. Wallis CJD, Butaney M, Satkunasivam R, Freedland SJ, Patel SP, Hamid O, et al. Association of Patient Sex With Efficacy of Immune Checkpoint Inhibitors and Overall Survival in Advanced Cancers: A Systematic Review and Meta-analysis. JAMA Oncol 2019 Jan 3 [cited 2019 Feb 18]; Available from: https://jamanetwork.com/journals/jamaoncology/fullarticle/2719757.
24. Yang F, Markovic SN, Molina JR, Halfdanarson TR, Pagliaro LC, Chintakuntlawar AV, et al. Association of Sex, Age, and Eastern Cooperative Oncology Group Performance Status With Survival Benefit of Cancer Immunotherapy in Randomized Clinical Trials: A Systematic Review and Meta-analysis. JAMA Netw Open. 2020;3(8):e2012534.
25. Terrin N, Schmid CH, Lau J, Olkin I. Adjusting for publication bias in the presence of heterogeneity. Stat Med. 2003;22(13):2113–26.
26. Harrer M, Cuijpers P, Furukawa TA, Ebert DD. Doing Meta-Analysis With R: A Hands-On Guide. 1st ed. Boca Raton, FL and London: Chapman & Hall/CRC Press; 2021. Available from: https://www.routledge.com/Doing-Meta-Analysis-with-R-A-Hands-On-Guide/Harrer-Cuijpers-Furukawa-Ebert/p/book/9780367610074.
27. Highfill SL, Cui Y, Giles AJ, Smith JP, Zhang H, Morse E, et al. Disruption of CXCR2-Mediated MDSC Tumor Trafficking Enhances Anti-PD1 Efficacy. Sci Transl Med. 2014;6(237):237ra67.
28. Contreras-Sandoval AM, Merino M, Vasquez M, Troconiz IF, Berraondo P, Garrido MJ. Correlation between anti-PD-L1 tumor concentrations and tumor-specific and nonspecific biomarkers in a melanoma mouse model. Oncotarget. 2016 Oct 18;7. Available from: https://doi.org/10.18632/oncotarget.12727.
29. van Elsas A, Hurwitz AA, Allison JP. Combination Immunotherapy of B16 Melanoma Using Anti-Cytotoxic T Lymphocyte-Associated Antigen 4 (Ctla-4) and Granulocyte/Macrophage Colony-Stimulating Factor (Gm-Csf)-Producing Vaccines Induces Rejection of Subcutaneous and Metastatic Tumors Accompanied by Autoimmune Depigmentation. J Exp Med. 1999;190(3):355.
30. Bonnotte B, Gough M, Phan V, Ahmed A, Chong H, Martin F, et al. Intradermal Injection, as Opposed to Subcutaneous Injection, Enhances Immunogenicity and Suppresses Tumorigenicity of Tumor Cells. Cancer Res. 2003;63(9):2145–9.
31. Thomas VA, Balthasar JP. Understanding Inter-Individual Variability in Monoclonal Antibody Disposition. Antibodies. 2019;8(4):56.
32. Audebert C, Laubreton D, Arpin C, Gandrillon O, Marvel J, Crauste F. Modeling and characterization of inter-individual variability in CD8 T cell responses in mice. In Silico Biol. 2020;14(1–2):13–39.
33. Karp NA, Wilson Z, Stalker E, Mooney L, Lazic SE, Zhang B, et al. A multi-batch design to deliver robust estimates of efficacy and reduce animal use - a syngeneic tumour case study. Sci Rep. 2020;10(1):6178.
34. Harrington DP, Fleming TR. A Class of Rank Test Procedures for Censored Survival Data. Biometrika. 1982;69(3):553–66.
35. Mestas J, Hughes CCW. Of Mice and Not Men: Differences between Mouse and Human Immunology. J Immunol. 2004;172(5):2731.
36. Schofield DJ, Percival-Alwyn J, Rytelewski M, Hood J, Rothstein R, Wetzel L, et al. Activity of murine surrogate antibodies for durvalumab and tremelimumab lacking effector function and the ability to deplete regulatory T cells in mouse models of cancer. mAbs. 2021;13(1):1857100.
37. Lechner MG, Karimi SS, Barry-Holson K, Angell TE, Murphy KA, Church CH, et al. Immunogenicity of murine solid tumor models as a defining feature of in vivo behavior and response to immunotherapy. J Immunother Hagerstown Md 1997. 2013;36(9):477–89.
38. Khalsa JK, Cheng N, Keegan J, Chaudry A, Driver J, Bi WL, et al. Immune phenotyping of diverse syngeneic murine brain tumors identifies immunologically distinct types. Nat Commun. 2020;11(1):3912.
39. Rizvi NA, Hellmann MD, Snyder A, Kvistborg P, Makarov V, Havel JJ, et al. Mutational landscape determines sensitivity to PD-1 blockade in non–small cell lung cancer. Science. 2015;348(6230):124.
40. Łuksza M, Riaz N, Makarov V, Balachandran VP, Hellmann MD, Solovyov A, et al. A neoantigen fitness model predicts tumour response to checkpoint blockade immunotherapy. Nature. 2017;8(551):517.
41. Daud AI, Loo K, Pauli ML, Sanchez-Rodriguez R, Sandoval PM, Taravati K, et al. Tumor immune profiling predicts response to anti–PD-1 therapy in human melanoma. J Clin Invest. 2016;126(9):3447–52.
42. Burtness B, Harrington KJ, Greil R, Soulières D, Tahara M, de Castro G, et al. Pembrolizumab alone or with chemotherapy versus cetuximab with chemotherapy for recurrent or metastatic squamous cell carcinoma of the head and neck (KEYNOTE-048): a randomised, open-label, phase 3 study. Lancet. 2019;394(10212):1915–28.
43. Ascierto PA, Del Vecchio M, Robert C, Mackiewicz A, Chiarion-Sileni V, Arance A, et al. Ipilimumab 10 mg/kg versus ipilimumab 3 mg/kg in patients with unresectable or metastatic melanoma: a randomised, double-blind, multicentre, phase 3 trial. Lancet Oncol. 2017;18(5):611–22.
44. Chen J, Wang J, Xu H. Comparison of atezolizumab, durvalumab, pembrolizumab, and nivolumab as first-line treatment in patients with extensive-stage small cell lung cancer: A systematic review and network meta-analysis. Medicine (Baltimore). 2021;100(15):e25180.
45. Huang Q, Zheng Y, Gao Z, Yuan L, Sun Y, Chen H. Comparative Efficacy and Safety of PD-1/PD-L1 Inhibitors for Patients with Solid Tumors: A Systematic Review and Bayesian Network Meta-analysis. J Cancer. 2021;12(4):1133–43.
46. Arce Vargas F, Furness AJS, Solomon I, Joshi K, Mekkaoui L, Lesko MH, et al. Fc-Optimized Anti-CD25 Depletes Tumor-Infiltrating Regulatory T

Cells and Synergizes with PD-1 Blockade to Eradicate Established Tumors. Immunity. 2017;46(4):577–86.
47. Selby MJ, Engelhardt JJ, Quigley M, Henning KA, Chen T, Srinivasan M, et al. Anti-CTLA-4 antibodies of IgG2a isotype enhance antitumor activity through reduction of intratumoral regulatory T cells. Cancer Immunol Res. 2013;1(1):32–42.

Publisher's Note