Estimating the replicability of highly cited clinical research (2004-2018)

Gabriel Gonçalves da Costa (ORCID: 0000-0002-1141-7827), Kleber Neves (ORCID: 0000-0001-9519-4909), Olavo B. Amaral (ORCID: 0000-0002-4299-8978)

Affiliation:

Institute of Medical Biochemistry Leopoldo de Meis, Federal University of Rio de Janeiro

Funding:

G.G.C. received funding from CNPq and CAPES. K.N. received funding from the Serrapilheira Institute. O.B.A. received funding from FAPERJ (E-26/200.824/2021), CNPq (308624/2018-1) and the Serrapilheira Institute.

Correspondence: Gabriel Gonçalves da Costa (gabriel.costa@bioqmed.ufrj.br)

Author Contributions:

Conceptualization: All authors

Methodology: All authors

Project Administration: All authors

Investigation (data collection): GGC and KN

Writing – Original Draft: GGC

Writing – Review and Editing: OBA and KN

Supervision: OBA

Conflicts of Interest:

None declared.

Abstract

Introduction: Previous studies about the replicability of clinical research based on the published literature have suggested that highly cited articles are often contradicted or found to have inflated effects. Nevertheless, there are no recent updates of such efforts, and this situation may have changed over time.

Methods: We searched the Web of Science database for articles studying medical interventions with more than 2000 citations, published between 2004 and 2018 in high-impact medical journals. We then searched for replications of these studies in PubMed using the PICO (Population, Intervention, Comparator and Outcome) framework. Replication success was evaluated by the presence of a statistically significant effect in the same direction and by overlap of the replication's effect size confidence interval (CIs) with that of the original study. Evidence of effect size inflation and potential predictors of replicability were also analyzed.

Results: A total of 89 eligible studies, of which 24 had valid replications (17 meta-analyses and 7 primary studies) were found. Of these, 21 (88%) had effect sizes with overlapping CIs. Of 15 highly cited studies with a statistically significant difference in the primary outcome, 13 (87%) had a significant effect in the replication as well. When both criteria were considered together, the replicability rate in our sample was of 20 out of 24 (83%). There was no evidence of systematic inflation in these highly cited studies, with a mean effect size ratio of 1.03 (95% CI [0.88, 1.21]) between initial and subsequent effects. Due to the small number of contradicted results, our analysis had low statistical power to detect predictors of replicability.

Conclusion: Although most studies did not have eligible replications, the replicability rate of highly cited clinical studies in our sample was higher than in previous estimates, with little evidence of systematic effect size inflation.

Introduction

The replicability of published research has been recently questioned in different scientific fields, with replication rates shown to be variable and often low (1–8). Whether this represents a "reproducibility crisis" is open to debate (9), and defining what constitutes a successful replication is not trivial (10). Systematic replication efforts have mostly focused on restricted samples of the literature, and data on the subject is still lacking in many areas.

The replicability of highly cited clinical research was studied by Ioannidis in 2005, based on available published replications of a sample of articles between 1990 and 2003 (11). It focused on the reproducibility of study conclusions, typically assessed by statistical significance, as well as on effect size comparisons. 44% of highly cited studies had been successfully replicated, 16% had been contradicted, 16% had found effects that were larger than those of subsequent studies, and 24% remained unchallenged.

A similar effort for highly cited psychiatry research between 2000 and 2002 found lower estimates, with 19% of studies replicated, 19% contradicted, 13% initially inflated and 48% unchallenged (12). Another study on critical care research found that 18% of interventions published in high-profile journals between 1946 and 2016 had their results replicated by a subsequent study, whereas 22% were contradicted, 2% had replications in progress and 58% remained unchallenged (13).

Clinical research has changed in some aspects over the last two decades. A priori registration of study protocols has become more common and mandatory for clinical trials in many countries (14). Although publication bias has not been eliminated (15), the likelihood of null results has increased in published studies (16). Reporting guidelines have become more widely used and underwent reevaluations and updates (17,18). The push for full reporting of results and availability of individual patient data has also gained ground (14,19). Thus, the replicability panorama in high-impact clinical research may have changed during this period (20–22).

In light of this, the goal of this study is to estimate the replicability of highly cited clinical studies published between 2004 and 2018. Our primary outcome is the rate of successful replication in these studies, as measured both by statistical significance in the same direction and by overlap of CIs for the main effect measure in both studies. We also explore effect size inflation and potential predictors of replicability.

Methods

An overview of the project, datasets and analysis code can be found at https://osf.io/a8zug/. The protocol for the study was preregistered at https://osf.io/nh965/, with a step-by-step methodology available at https://osf.io/2qncz/ and updates and amendments described at https://osf.io/26d98/. All statistical analyses were performed in R version 4.1.2 (23). Data and analysis code are available at https://osf.io/5qhdz and https://osf.io/9hmx4, respectively.

Search for highly cited studies

We searched the Web of Science database for articles with more than 2000 citations published between January 1st, 2004, and December 31st, 2018, in medical journals with an impact factor above 14 in the 2020 Journal Citation Reports (list at https://osf.io/2qncz/). General journals were searched on February 7th, 2020, and specialty journals on March 4th, 2020. The cutoffs for citation and impact factors were twice as large as those used by Ioannidis (11), accounting for the growth in the total number of articles in PubMed during the period (calculation at https://osf.io/t9xu7/).

Within this sample, one author (K.N.) screened titles and abstracts for articles that addressed the efficacy of therapeutic or preventive interventions with primary data (i.e., excluding reviews, meta-analyses or articles that combined two or more previous studies). Two evaluators (G.G.C. and K.N.) then selected the primary outcome in each study, or the main conclusion in the abstract if the study had no primary outcome. In the case of co-primary outcomes or equally emphasized conclusions (11,24–26), we chose the

outcome that was deemed more clinically relevant (e.g., mortality over progression or neurologic improvement over reperfusion). In the case of trials with more than two arms (27), we selected the most effective drug as the intervention and randomly chose an active comparator. Studies with no control group (e.g., phase 1 trials) were considered eligible if the abstract clearly stated that an intervention was clinically effective. When both benefits and harms or caveats were presented, focus was given on the net conclusion of whether the experimental intervention merited consideration for use in clinical practice. Disagreements in outcome selection were solved by consensus with the help of a third investigator (O.B.A.).

For each article, the study design, sample size, journal name and category (general or specialty) were extracted. We also extracted the selected outcome measure – i.e., odds ratio (OR), relative risk (RR), hazard ratio (HR), incidence rate ratio (IRR) or objective response rate (ORR) – with its effect size and respective CI. For controlled studies, results were classified as positive or negative according to the authors' stated statistical significance threshold. Non-inferiority trials were classified as positive only if the intervention was found to be superior (i.e., not merely non-inferior) to the comparator.

For each result, the population, intervention, comparator and outcome (PICO) (28), both in specific (e.g., "myocardial infarction, ischemic stroke, unstable angina, or cardiovascular surgery") and general forms (e.g., "cardiovascular events") (14) were extracted. Two evaluators (K.N. and G.G.C.) described PICO components independently and resolved disagreements by consensus. Results of the independent extraction and consensus decisions can be found at https://osf.io/sfdxv.

Search for replications

After agreement was reached on PICO components, two evaluators (G.G.C. and K.N.) performed independent searches for replications of highly cited studies in PubMed. Search terms were defined independently by each evaluator and included the name of the drug or intervention, the general form of the outcome, and the population (i.e., clinical condition) as described in the article's title, along with corresponding Medical Subject Headings (MeSH)

terms. The comparator was included in the search strategy only if it was an active intervention (i.e., not a placebo or sham). Details can be found at https://osf.io/zv65u/.

A study was considered a replication of the highly cited study when it shared the same PICO general components, namely (a) the drug or intervention, without considering dose or regimen (except for studies performing dose or regimen comparisons), (b) the general form of the outcome (as described in (14)), (c) the population/clinical condition as described in the highly cited article's title and (d) the comparator. When geographical information was included as a descriptor of the population (e.g., "European patients") (29,30), we did not include this information as part of the population component (31–33).

Replications needed to be (a) a study type with higher strength of evidence (34) (i.e., randomized controlled trials (RCTs) over cohort studies over smaller uncontrolled studies) or (b) a similar study type with a sample size equal to or larger than the original study. Meta-analyses were considered as eligible replications if the highly cited study accounted for less than half of their sample size. For network meta-analyses, only the sample size of the direct comparison between the intervention and comparator counted for this purpose. If a meta-analysis (35,36) included a single additional study beyond the highly cited one (32,37), we considered the effect size of this study as the replication (31,38), rather than that of the entire meta-analysis. If more than one replication was found, the one with the largest sample size for the specific comparison was considered.

When different replications were selected by each evaluator, both were made available for the two evaluators to choose the best option independently. Disagreements in this step were solved by consensus with the participation of a third author (O.B.A). Agreement in the initial selection was 36%, but rose to 91% when selected replications were made available to both evaluators. Agreement data can be found at https://osf.io/qz6u9, with resolution of disagreements detailed at https://osf.io/ma9bn.

After identifying the best available replication, both evaluators independently selected the outcome and effect size from the replication that corresponded most closely to the one in the original study. Disagreements were solved by consensus. For network meta-analyses,

direct comparisons were favored over indirect ones when both were available, either in the manuscript or supplementary material. Agreement data for this process can be found at https://osf.io/2c6jx. Changes in the choice of replication and effect size during analysis are documented at https://osf.io/jq7ec.

As the effect estimates of meta-analyses usually included the highly cited study and were thus not fully independent from it, we re-estimated these effects after removing the highly cited when enough information was provided for this purpose. For this, we used the primary study results as retrieved from the meta-analysis, estimating effect sizes based on numbers of events and patients when these were available, or the log-transformed point estimate of the RR or HR when they were not, with a standard error estimated by $\frac{[\log(\text{upper limit of CI}) - \log(\text{lower limit of CI})]}{3.92}$. For data synthesis, a random-effects model was performed using the Mantel-Haenszel method for effect size estimation in the package meta in the R software for statistical computing (39). Replicability rates using the effect sizes from these fully independent meta-analyses are provided in addition to the main results as a supplementary analysis.

Evaluating Replication Success

Pairs of highly cited studies and their replications were analyzed to evaluate whether results were successfully replicated on the basis of two criteria: (a) statistical significance (an effect in the replication with $p < 0.05$ in the same direction as that observed in the highly cited study) and (b) confidence interval overlap (an overlap of the 95% CIs for the outcome of interest in both studies). When the highly cited study presented a non-significant effect or did not include a statistical comparison (e.g., phase 1 trials), only the second criterion was used. The primary outcome was the rate of successful replication in our sample by both criteria (or by CI overlap alone when statistical significance was not applicable). As additional criteria, we analyzed whether the replication point estimate was contained in the 95% CI of the highly cited study and vice versa. A sensitivity analysis was performed applying the statistical significance criterion to initially non-significant studies as well. In

one case where the replication was a Bayesian meta-analysis, the original study's CI was compared to a credible interval (CrI), in the case of minimally informative priors (40). P-values were calculated from effect sizes (point estimates and confidence intervals) for each replication and highly cited study (details at https://osf.io/jbn83).

When outcome measures differed between highly cited studies and replications (e.g., RR in the highly cited study vs. OR in the replication or vice-versa) and the replication was a primary study, the replication measure was converted to the one in the highly cited study using the data available in the article. When the replication was a meta-analysis that included the highly cited study, we chose the risk measure that was used for data synthesis, using the original study's effect size as included in the meta-analysis (details at https://osf.io/rfqgd). If the highly cited study was not included in the meta-analysis (e.g., when the meta-analysis included an update of the highly-cited study with a longer follow-up), we manually converted the outcome measure of the highly cited study to the one in the meta-analysis using the original data. When the original study was a phase 1 trial measuring ORR, we manually calculated this measure in replications when needed, with CIs based on the Clopper-Pearson exact method. In the meta-analysis by Hamid et al. (41), ORRs for both RCTs that were eligible replications of the highly cited study (a phase 1 trial) were calculated manually based on the combined data. Details can be found at https://osf.io/mfwv2.

95% CIs for replicability rates were calculated by $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$. Replicability rates in the main results use effect sizes from published replications, while supplementary analyses use only fully independent replications, recalculating meta-analytic effect sizes in the absence of the highly cited study (and excluding meta-analyses for which this was not possible).

Effect size inflation

Effect size inflation was estimated on the basis of ratios between the effect sizes of highly cited studies and replications. For unfavorable outcomes (e.g., death, tumor progression), in which effectiveness increases as the outcome measure decreases, the inflation ratio was

defined as the point estimate of the replication divided by that of the original study. For favorable outcomes (e.g., neurologic improvement), in which effectiveness increases along with the outcome measure, it was defined as the point estimate of the original study divided by that of the replication. Publication order was considered in this calculation: thus, when the replication was a meta-analysis in which the pooled sample size of studies preceding the highly cited study was larger than that of those that followed it, we inverted the ratios, considering the highly cited study as the replication and vice-versa for this purpose. This was also performed in a case where the replication was an RCT (identified within a meta-analysis (37) that was published before the highly cited one (31). CIs for mean effect size inflation were calculated by the Wilson score interval.

As these two adjustments had not been pre-specified in the protocol, we performed sensitivity analyses using different ways to deal with positive/negative outcomes and study order within meta-analyses when analyzing effect size inflation. For the former, coining of the effects was performed to convert all favorable outcomes to unfavorable outcomes: overall response rates were subtracted from 1 (1 – ORR), odds ratios were inverted (1/OR) and relative risks for the complementary outcome were calculated based on the original data (https://osf.io/5tmus and https://osf.io/7wtr8). For the latter, we analyzed data considering the highly cited study as the original one, independent of study order in the meta-analysis. As done for replication rates, we also provide supplementary effect size inflation analyses based on fully independent replications only, using recalculated meta-analytical estimates in the absence of the highly cited study.

For analysis of effect size inflation, natural logarithms of the ratios were used for each study pair (including those with initially negative results) to calculate the mean and CI of these ratios, both for the whole sample and for phase 1 trials and RCTs separately. For the whole sample, we performed a one-sample t-test against a theoretical mean of 0, which would indicate absence of systematic inflation. Although these calculations were performed using log-transformed values to correct for the inherent asymmetry of ratios, we transformed means and CIs back to a linear scale for clarity when describing results.

Predictors of Replicability

Finally, we analyzed if studies with contradicted results – i.e., those failing in one or both replication criteria – differed from successfully replicated ones in the following aspects: (a) study design (RCTs vs. other designs); (b) nature of intervention, (pharmacological vs. non-pharmacological); (c) sample size; (d) p-value of the original study; and (e) citations per year. To compare these aspects between replicated and contradicted studies, we used Fisher's exact test for categorical variables (a and b) and Mann-Whitney's U test for continuous variables (c through e). We had also planned to use the effect size of the highly cited study as a predictor, but due to the heterogeneity in outcome measures, which included both proportions (i.e., overall response rates) and measures of association (i.e., ORs, RRs, IRRs and HRs) converting them to a single effect size measure turned out to be unfeasible.

Results

Results from our systematic search of the literature are shown as a flowchart in **Figure 1**. A total of 89 highly cited studies met our inclusion criteria. Of these, 24 had an eligible replication according to our criteria.

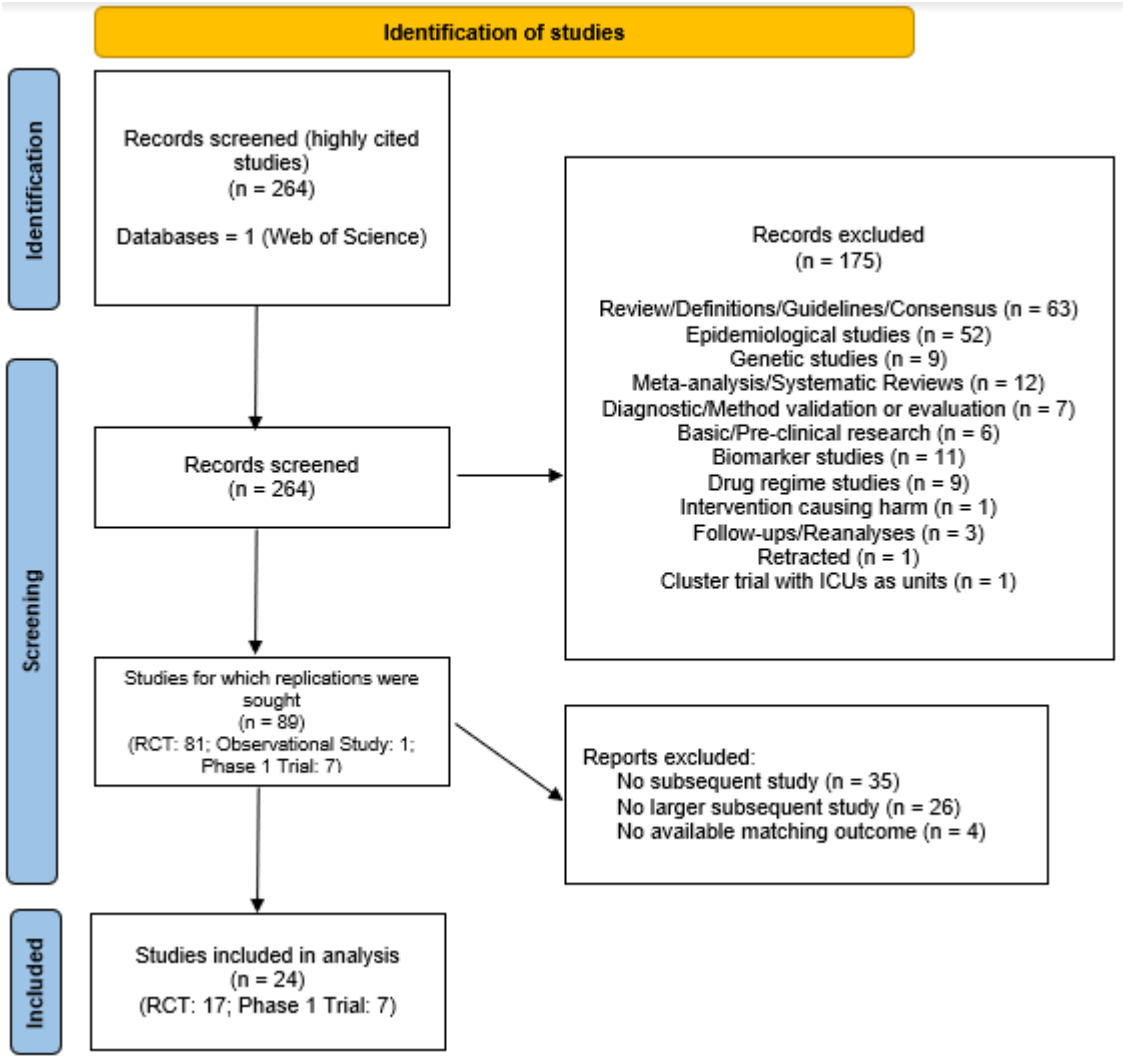

Figure 1. PRISMA flowchart. 89 eligible highly cited studies were found, of which 24 had an eligible replication. A complete list of the studies can be found at https://osf.io/ub38r/. A more detailed list of reasons for inclusion/exclusion is available at https://osf.io/ma9bn/.

As shown in **Table 1,** included studies received a median of 2842 citations, and were mostly RCTs of pharmacological interventions in cancer or heart disease, with some phase 1 cancer trials as well.

Description	Total (n = 89)	%	Replication found (n = 24)	%

Description		Total (n = 89)	%	Replication found (n = 24)	%
Highly cited study design	RCT	81	91%	17	71%
	Phase 1 trial	7	8%	7	29%
	Cohort study	1	1%	0	0%
Type of intervention	Pharmacological	72	81%	16	67%
	Non-pharmacological	17	19%	8	33%
Journal	*New England Journal of Medicine*	80	90%	22	92%
	Lancet	4	4%	0	0%
	Lancet Oncology	4	4%	2	8%
	JAMA	1	1%	0	0%
Condition	Cancer	53	60%	15	63%
	Cardiovascular	27	30%	8	33%
	Other	9	10%	1	4%
Citations	< 3,000	45	51%	13	54%
	3,000 - 4,000	26	29%	8	33%
	4,000 - 6,000	12	13%	1	4%
	> 6,000	6	7%	2	8%

Table 1. Features of highly cited studies. Columns show the numbers of highly cited studies and of those for which an eligible replication was found in each category. Percentages refer to the total number of highly cited studies (n=89) or studies with replications (n=24), respectively.

Most replications were direct-comparison meta-analyses, followed by RCTs, network meta-analyses and a phase 2 trial (**Table 2**), with RCTs more commonly representing replications of phase 1 trials. Two meta-analyses (42,43) replicated more than one highly-cited study in the sample (2 each). All phase 1 trials had available replications in the literature, while the only cohort study in our sample had no eligible replication.

		Replication study design			
Design of highly cited study	**Replication found (%)**	**Phase 2 Trial**	**RCT**	**Meta-analysis**	**Network meta-analysis**
Cohort study	0/1 (0%)	0	0	0	0
Phase 1 trial	7/7 (100%)	1	3	2	1
RCT	17/81 (21%)	0	3	10	4
Total	24/89 (27%)	1	6	12	5

Table 2. Features of replications. Percentages refer to the total number of highly cited studies in each category. For studies with a replication, the table describes the type of study; most studies, however, did not have a valid replication. A network meta-analysis including a direct comparison for the groups of interest (44) was classified as a regular meta-analysis. Two meta-analyses (42,43) replicating two highly cited studies are counted twice in their respective categories.

A list of claims from highly cited studies is shown in **Table 3**. With few exceptions, most of them made claims of efficacy in their abstracts. Efficacy in phase 1 trials was measured by ORR (n=7), while differences in outcomes in RCTs were measured by HR (n=8), RR (n=5), IRR (n =1) or OR (n=3). All phase 1 trials made clear claims of efficacy based on tumor regression. Among RCTs, 2 were negative, with a p-value above the standard cutoff of 0.05.

Highly Cited Study	Replication	PICO General Components	Original conclusion	Effect Measure	Highly Cited ES [95% CI]	Replication ES [95%CI]
Brahmer et al., 2012 (45)	Zhang et al., 2016 (46)	Advanced cancer, Nivolumab, NA, Response	Nivolumab induced tumor regression and prolonged stabilization of disease in advanced cancers	ORR	0.13 [0.06-0.17]	0.27 [0.21-0.33]
Topalian. et al. 2012 (47)	Tie et al., 2017 (48)	Cancer, Nivolumab, NA, Tumour Response	Nivolumab produced objective responses in cancer regression.	ORR	0.21 [0.16-0.26]	0.26 [0.21-0.31]
TAXUS-IV (Stone et al., 2004) (49)	Bangalore et al., 2013 (50)	Coronary artery disease, Paclitaxel stent, Metal stent, Revascularization	A paclitaxel-eluting stent reduced the rate of clinical and angiographic restenosis at nine months	RR	0.39 [0.26-0.59]	0.66 [0.59-0.74]
SYNTAX (Serruys et al., 2009) (51)	Ali et al., 2018 (52)	Severe coronary artery disease, Percutaneous coronary intervention (PCI), Coronary-artery bypass grafting (CABG), Cardiovascular and cerebrovascular events	CABG resulted in lower rates of major adverse cardiac or cerebrovascular events than PCI after 1 year	OR	1.44 [1.11-1.89]	1.42 [1.27-1.59]
HERA (Piccart-Gebhart et al., 2005) (53)	Genuino et al., 2019 (54)	HER2-positive breast cancer after adjuvant chemotherapy, Trastuzumab, Observation, Progression	Trastuzumab after adjuvant chemotherapy improved disease-free survival in HER2-positive breast cancer	HR	0.54 [0.43-0.67]	0.65 [0.55-0.75]

Highly Cited Study	Replication	PICO General Components	Original conclusion	Effect Measure	Highly Cited ES [95% CI]	Replication ES [95%CI]
CATIE (Lieberman et al., 2005) (27)	Soares-Weiser et al., 2013 (55)	Schizophrenia, Olanzapine, Quetiapine, Treatment discontinuation	Time to discontinuation of treatment for any cause was longer for olanzapine than for quetiapine	HR	0.63 [0.52-0.76]	0.68 [0.56-0.83]
SHARP (Llovet et al., 2008) (56)	Niu et al., 2016 (42)	Advanced hepatocellular carcinoma, Sorafenib, Placebo, Survival	Survival was longer with sorafenib than with placebo	HR	0.69 [0.55-0.87]	0.69 [0.60-0.79]
ERSPC (Schröder et al., 2009) (29)	Ilic et al., 2018 (57)	Middle- to old-age men, PSA screening, Control (no screening), Prostate cancer death	Periodic PSA-based screening reduced the rate of death from prostate cancer after a median follow-up of 9 years	IRR	0.80 [0.65-0.98]	0.96 [0.85-1.08]
PROFILE 1007 (Shaw et al., 2013) (58)	Elliott et al., 2020 (59)	Advanced ALK-positive lung cancer, Crizotinib, Chemotherapy, Progression	Crizotinib was superior to standard chemotherapy in preventing cancer progression	HR	0.49 [0.37-0.64]	0.46 [0.39-0.54]
ACCORD (Action to Control Cardiovascular Risk in Diabetes Study Group et al., 2008) (60)	Fang et al., 2016 (61)	Type 2 diabetes, Intensive glucose control, Standard therapy, Cardiovascular events or cardiovascular death	As compared with standard therapy, intensive therapy did not reduce major cardiovascular events after a mean follow-up of 3.5 years	RR	0.95 [0.82-1.09]	0.92 [0.85-1.00]

Highly Cited Study	Replication	PICO General Components	Original conclusion	Effect Measure	Highly Cited ES [95% CI]	Replication ES [95%CI]
Cheng et al., 2009 (33)	Niu et al., 2016 (42)	Advanced hepatocellular carcinoma, Sorafenib, Placebo, Survival	Sorafenib increased survival in advanced hepatocellular carcinoma when compared with placebo	HR	0.68 [0.50-0.93]	0.69 [0.60-0.79]
EXTEND-IA (Campbell et al., 2015) (26)	Goyal et al., 2016 (62)	Ischemic stroke, Endovascular thrombectomy + Alteplase, Alteplase, Disability	Early thrombectomy with the Solitaire FR stent retriever improved disability in ischemic stroke as compared with alteplase alone	OR	6 [2-18]	4.04 [2.75-5.93]
PARTNER A (Smith et al., 2011) (35)	US Core Valve Study (Adams et al.) (38)	Aortic stenosis, Transcatheter aortic-valve implantation (TAVI), Surgical replacement, Mortality	Transcatheter and surgical procedures for aortic valve replacement were associated with similar rates of survival at 1 year in high-risk patients	RR	0.98 [0.75-1.26]	0.73 [0.54-0.98]
MR CLEAN (Berkhemer et al., 2015) (63)	Rodrigues et al., 2016 (43)	Acute ischemic stroke, Intraarterial treatment + Usual care, Usual care, Disability	Intraarterial treatment administered within 6 hours after stroke onset was effective in reducing disability assessed at 90 days post-intervention.	RR	1.73 [1.27-2.35]	1.37 [1.14-1.64]
ECASS III (Hacke et al., 2008) (64)	Wardlaw et al., 2012 (65)	Acute ischemic stroke, Alteplase, Placebo, Disability	Intravenous alteplase improved disability in patients with acute ischemic stroke assessed 90 days after the intervention	OR	1.34 [1.02-1.76]	1.29 [1.16-1.43]

Highly Cited Study	Replication	PICO General Components	Original conclusion	Effect Measure	Highly Cited ES [95% CI]	Replication ES [95%CI]
EURTAC (Rosell et al., 2012) (30)	Zhao et al., 2019 (44)	Advanced EGFR mutation-positive non-small-cell lung cancer, Erlotinib, Standard chemotherapy, Progression	Erlotinib increased progression-free survival when compared to standard chemotherapy in Asian and European patients with advanced EGFR mutation-positive non-small-cell-lung cancer	HR	0.37 [0.25-0.54]	0.23 [0.17-0.30]
ESCAPE (Goyal et al., 2015) (66)	Rodrigues et al., 2016 (43)	Ischemic stroke, Standard care + Endovascular treatment, Standard care, Disability	Rapid endovascular treatment (thrombectomy) improved functional outcomes in up to 12 hours after symptom onset	RR	1.86 [1.39-2.47]	1.37 [1.14-1.64]
NEJSG (Maemondo et al., 2010) (32)	IPASS (Mok et al., 2009) (31)	Non-small-cell lung cancer with mutated EGFR, Gefitinib, Carboplatin–paclitaxel, Progression	First-line gefitinib improved progression-free survival as compared with standard chemotherapy in patients with non-small-cell lung cancer with EGFR mutations.	HR	0.32 [0.22-0.41]	0.48 [0.36 – 0.64]
Hamid et al., 2013 (41)	Pyo et al., 2017 (67)	Melanoma, Lambrolizumab, NA, Tumor response	In advanced melanoma, treatment with lambrolizumab resulted in a high rate of sustained tumor regression	ORR	0.38 [0.25-0.44]	0.29 [0.26-0.32]
KEYNOTE-024 (Reck et al., 2016) (68)	KEYNOTE-042 (Mok et al., 2019) (69)	PD-L1-positive non-small-cell lung cancer, Pembrolizumab, Chemotherapy, Survival	Pembrolizumab led to longer overall survival than platinum-based chemotherapy for PD-L1-positive non-small-cell lung cancer	HR	0.50 [0.37-0.68]	1.07 [0.94-1.21]
KEYNOTE-001 (Garon et al., 2015) (70)	KEYNOTE-010 (Herbst et al., 2016) (71)	Non–small-cell lung cancer, Pembrolizumab, NA, Tumor Response	Pembrolizumab showed antitumor activity and led to objective responses in patients with advanced non–small-cell lung cancer	ORR	0.19 [0.16-0.23]	0.25 [0.22-0.29]

Highly Cited Study	Replication	PICO General Components	Original conclusion	Effect Measure	Highly Cited ES [95% CI]	Replication ES [95%CI]
Wolchock et al., 2013 (72)	Checkmate 067 (Larkin et al., 2015) (25)	Advanced Melanoma, Nivolumab + Ipilimumab, NA, Response	Concurrent therapy with nivolumab and ipilimumab led to tumor regression in a substantial proportion of patients	ORR	0.40 [0.27-0.55]	0.44 [0.38-0.49]
Flaherty et al., 2010 (73)	BRIM-3 (Chapman et al., 2011) (24)	Metastatic melanoma with activated BRAF, PLX4032 (Vemurafenib), NA, Response	Treatment of metastatic melanoma carrying the V600E BRAF mutation with PLX4032 (Vemurafenib) resulted in tumor regression	ORR	0.81 [0.63-0.93]	0.48 [0.42-0.55]
Fong et al., 2009 (74)	Kaufman et al., 2015 (75)	BRCA1- or BRCA2-mutated resistant cancers, Olaparib (AZD2281), NA, Tumor Response	Olaparib had antitumor activity and led to objective responses in cancers associated with BRCA1 or BRCA2 mutations	ORR	0.47 [0.24-0.71]	0.26 [0.21-0.32]

Table 3. Summary of the 24 highly cited studies with replications. Table shows the references for both studies, the general PICO components, the original study's conclusion, the outcome measure and the effect sizes (with CIs) found in the original study and replication. Replication studies for PARTNER A (35) and NEJSG (32) were obtained from 2-study meta-analyses found in our search (36,37) and their effect sizes are drawn from these meta-analyses; in the case of IPASS (31), this corresponds to a subgroup matching that of the highly cited study. Two pairs of studies (MR CLEAN (63) and ESCAPE, Cheng et al. 2009 (33) and SHARP (56)) are replicated by the same meta-analyses (Rodrigues et al. 2016 (43) and Niu et al. 2016 (42)).

When a meta-analysis was selected as a replication, the highly cited study could come after some (or most) of the studies in the meta-analysis, and thus consist of a replication of previous literature itself. **Figure 2**, shows the relative sample sizes of the highly cited studies in their replications. On average, highly cited studies corresponded to 18% of the total sample size of the meta-analyses, but this number ranged from 2% to 42% (samples sizes can be found at https://osf.io/ma9bn). Most meta-analyses had a larger number of patients after the highly cited study than before it, with three exceptions (30,33,66), including one (42) in which all other studies in the meta-analysis preceded the highly cited one (33). In another case, a meta-analysis of 2 studies (37) led to an RCT published before the highly cited study (31) to be selected as a replication.

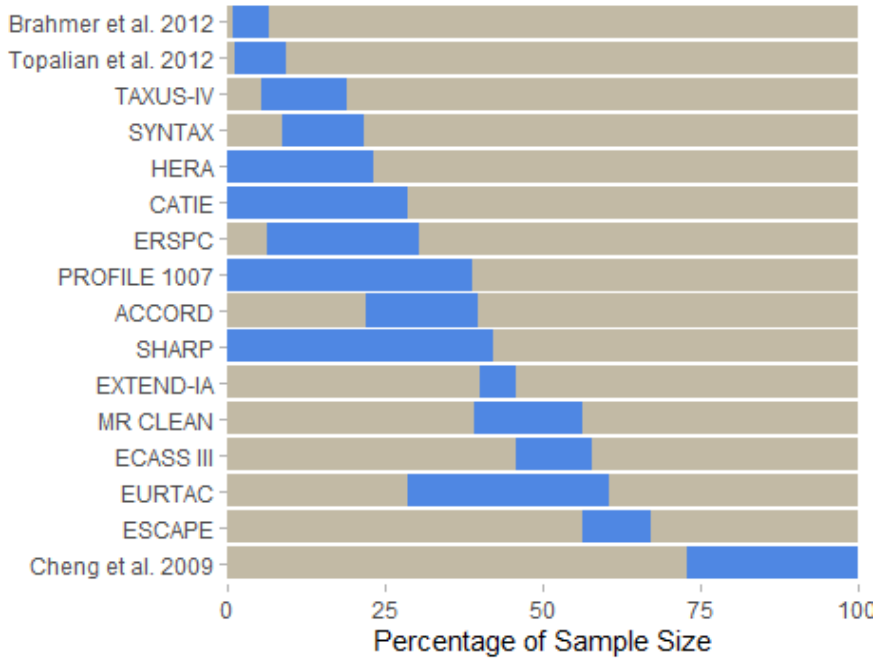

Figure 2. Relative contribution of highly cited studies to the sample size of replication meta-analyses. For studies with meta-analyses as replications, blue bars show the fraction of the total sample size that corresponds to the highly cited study, while gray bars represent the rest of the sample size. Fractions are arranged in temporal order from left to right, so that gray bars to the left of the blue ones represent studies that came before the highly cited one. The meta-analysis (67) replicating Hamid et al., 2013 (41) is not shown because the highly cited study is not included in it due to the lack of a control group.

Replication rates using different criteria are shown in **Table 4**. Among the 15 highly cited studies with statistically significant results, only 2 (13%) had a non-significant

result in the replication, whereas the 2 highly cited studies with negative results had significant results in their replications (albeit marginally so). We did not consider the latter as replication failures in our main analysis, as lack of significance in null hypothesis tests should not be taken as evidence of equivalence, especially considering that sample size was higher in the replications. However, we did perform a sensitivity analysis using the statistical significance criterion for studies with non-significant results as well.

Phase 1 trials with no control group were also not considered for the statistical significance criterion. That said, among the 7 phase 1 trials, 6 had replications with statistically significant results when comparing the intervention to a control group in a RCT; the remaining one was replicated by an uncontrolled phase 2 trial (75).

Concerning the effect size criterion, 21 out of 24 studies (88%) had overlapping 95% CIs, with 2 phase 1 trials and 1 RCT failing this criterion. In total, 15 out of 24 replications (62%) had point estimates that were contained in the CIs of the highly cited studies; conversely, only 9 (38%) of the original point estimates were included in the replication's CIs. That said, as CIs get narrower with increasing sample size, the latter criterion is excessively strict and should not be considered as a measure of replicability.

Criterion	Total	Replicated	% Replicated [95% CI]
Statistical significance	15	13	87% [62, 96]
95% CI overlap	24	21	88% [69, 96]
95% CI overlap and statistical significance	24	20	83% [64, 93]
Statistical significance (including negative studies)	17	13	76% [56; 96]
Replication estimate within highly cited study's 95% CI	24	15	62% [43, 79]
Highly cited study estimate within replication 95% CI	24	9	38% [21, 57]

Table 4. Rate of successful replication as measured by different criteria. Results are shown as the percentage of studies with eligible replications in which the original result was replicated by each of 5 different criteria. For the statistical significance criterion, we excluded 7 phase 1 trials that did not report p-values and 2 studies with non-significant results. For these studies, the aggregate criterion (CI overlap +

statistical significance) only considers CI overlap. The 2 negative studies are included in the "statistical significance (including negative studies)" criterion as a sensitivity analysis.

One major limitation of this analysis is that effect sizes from meta-analyses are not fully independent from the highly cited study when it is included in the estimate. To circumvent this, we recalculated meta-analytical effect estimates in the absence of the highly cited studies (**Table S1**). This was technically feasible without re-extracting data from the original studies for 10 out of 14 meta-analyses (replicating a total of 11 highly cited studies). The remainder were network meta-analyses without direct comparisons (42,50,59,67) or individual patient-data meta-analyses (62).

When using only fully independent replications (i.e., excluding meta-analyses that could not be reanalyzed), replicability rates were 80% for the statistical significance criterion, 84% for the CI overlap and 79% for the aggregated criterion (**Table S2)**. Differences between this and the main analysis were due to the different samples used in each of them, as no meta-analysis changed its replication status in either criterion when reanalyzed without the highly-cited study.

No evidence of effect size inflation was observed in our sample (**Figure 3** and **Table 5**), with the average ratio between the effect sizes of replications and those of highly cited studies approaching 1 when publication order was considered. Inflation increased slightly when effect sizes were coined (i.e., when the percentage of non-responders was used as the outcome) and when publication order was not taken into account (i.e., when highly cited studies were always considered as the reference), but remained low on average and did not reach statistical significance in any of our analyses. This picture did not change when only fully independent replications were used to estimate inflation (**Figure S1** and **Table S3**).

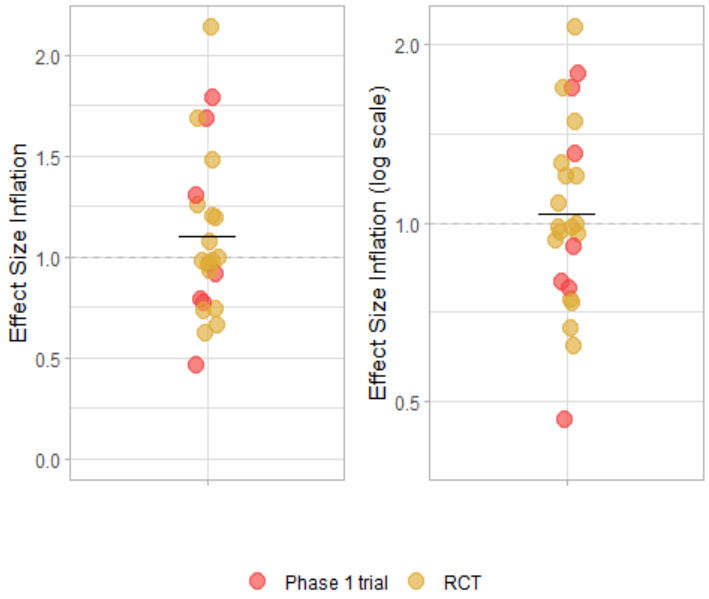

Figure 3. Effect size inflation. Effect size inflation was calculated for each study pair considering the effect sizes of highly cited studies and their respective replications. For unfavorable outcomes (e.g., death, tumor response), inflation was calculated as the ratio between the replication and highly cited point estimates. For favorable outcomes (e.g., neurologic improvement), it was calculated as the ratio between the highly cited and the replication point estimates. When the highly cited study came after most or all of the replication studies, we considered the highly cited study as the replication and inverted the ratio. Left panel shows the results in a linear scale (in which distribution is expected to be skewed upwards even in the absence of inflation because of the nature of ratios), while right panel shows them in a log-scale (in which values should be distributed symmetrically around 1 in the absence of inflation). Lines indicate the mean of the plotted values (which in the left panel differs from that on **Table 5**, calculated on the basis of log-transformed values). Colors indicate Phase 1 trials (red) and RCTs (orange).

Analysis	All	p-value	Phase 1 trials	RCTs
Main analysis	1.03 [0.88 - 1.21]	0.68	1.01 [0.64 – 1.57]	1.04 [0.88 - 1.23]
Publication order	1.10 [0.95 - 1.28]	0.20	1.01 [0.64 – 1.57]	1.14 [0.98 - 1.33]
Effect coining	1.07 [0.93 - 1.24]	0.34	1.17 [0.80 – 1.71]	1.03 [0.88 - 1.22]
Publication order + coining	1.14 [0.99 - 1.30]	0.07	1.17 [0.80 – 1.71]	1.12 [0.96 - 1.31]

Table 5. Analyses of effect size inflation. Analyses to detect evidence of effect size inflation using different ways to handle favorable/unfavorable outcomes and meta-analyses with most of the sample size preceding the highly cited study. In "Publication order", we did not invert the ratio if a replication had most or all of its sample size published before the highly cited study. In "effect coining", we used

unfavorable outcomes for all effect measures, thus avoiding the need to reverse ratios for favorable ones. The bottom row combines both approaches. P-values refer to a one-sample Student's t test against a theoretical mean of 0 for the log-transformed inflation ratios for each study pair. Values for means and 95% CIs were calculated for these log-transformed ratios and exponentiated back to the original scale. Inflation point estimates were manually replicated with the results of the R code output (https://osf.io/9hmx4) and this analysis can be found at https://osf.io/qnfgh.

Potential predictors of replicability are shown in **Table 6**. Although this analysis was planned in the protocol, it has low statistical power due to the low number of contradicted studies in our sample. The same analysis using only fully independent replications is shown on **Table S4.** In both analyses, low power prevents us from drawing any definite conclusions on predictors of replicability.

Predictor	Criteria	Median [IQR] Replicated	Median [IQR] Contradicted	p-value
Citations/year of the highly cited study	Significance [p < 0.05]	292 [250 – 454]	441 [338 – 543]	0.69
	CI overlap	329 [250 – 454]	494 [368 – 570]	0.35
	Significance [p < 0.05] & CI overlap	337 [270 – 459]	368 [241 – 532]	0.79
p-value of the highly cited study	Significance [p < 0.05]	$3x10^{-5}$ [$7x10^{-7}$ – $2x10^{-3}$]	0.02 [$8x10^{-3}$ – 0.02]	0.48
	CI overlap	$1x10^{-3}$ [$4x10^{-6}$ – 0.02]	$1x10^{-5}$ [$1x10^{-5}$ – $1x10^{-5}$]	0.82
	Significance [p < 0.05] & CI overlap	$5x10^{-4}$ [$3x10^{-6}$ – 0.01]	0.02 [$8x10^{-3}$ – 0.02]	0.72
Sample size of the highly cited study	Significance [p < 0.05]	551 [294 – 944]	182000 [182000 – 182000]	0.15
	CI overlap	602 [306 – 1557]	207 [207 – 207]	0.38
	Significance [p < 0.05] & CI overlap	551 [301 – 1191]	91104 [45655 – 136552]	0.82

Predictor	Criteria	# replicated by study design	# not replicated by study design	p-value
Highly cited study design	CI overlap	Phase 1 trial: 5/7 RCT: 16/17	Phase 1 trial: 2/7 RCT: 1/17	0.19
	Significance [p < 0.05] & CI overlap	Phase 1 trial: 5/7 RCT: 15/17	Phase 1 trial: 2/7 RCT: 2/17	0.55
	Significance [p < 0.05]	Pharmacological: 7/8 Other: 6/7	Pharmacological: 1/8 Other: 1/7	1.00
Type of intervention	CI overlap	Pharmacological: 13/16 Other: 8/8	Pharmacological: 3/16 Other: 0/8	0.53
	Significance [p < 0.05] & CI overlap	Pharmacological: 13/16 Other: 7/8	Pharmacological: 3/16 Other: 1/8	1.00

Table 6. Predictors of replication success. Table shows comparisons of potential predictors of replicability between replicated and contradicted studies using different definitions of replication success. P-values are shown for a Mann-Whitney test (for continuous variables) or Fisher's exact test (for categorical variables) comparing replicated and contradicted studies. For the significance criterion, only 15 studies (i.e., initially positive RCTs) were considered. Study design is not included as a predictor of statistical significance, as phase 1 trials did not use significance testing. Sample size for each group varies according to the specific criteria: statistical significance: 13 replicated, 2 contradicted; CI overlap: 21 replicated, 3 contradicted; both criteria: 20 replicated, 4 contradicted.

Discussion

Replicability of the literature

The replicability of highly cited clinical studies in our sample was high, with an 83% replication rate when considering our primary outcome of overlapping CIs along with statistical significance in the same direction. Using only fully independent replications (i.e., including meta-analyses only when highly-cited studies were removed) led to a slightly lower but still reasonably high estimate of 79%. Moreover, we did not find evidence of systematic effect size inflation, either for phase 1 trials or for RCTs.

These replication success rates are higher than those found in previous studies of highly cited clinical literature, where rates of 59% in general medicine (11) and 37% in psychiatry (12) were described when both statistical significance and effect size were considered, although the criteria for comparing effect sizes differed in each study. For statistical significance alone (a more homogeneous criterion), the successful replication

rate for studies with significant results was 87% in our study, as compared to 79% (11) and 63% (12) in the previous two studies, respectively.

Although the differences in these estimates could be due to changes in the replicability of the published literature, methodological discrepancies between studies should be considered. Ioannidis' study (11), used a broader definition of replication; thus, many article pairs in his sample would not have been considered to have matching PICO components by our criteria. Moreover, among the contradicted or inflated studies in his sample, 4 were cohort studies whose replications were RCTs or meta-analyses of RCTs. The author himself acknowledges that it is not always possible to validate exposures as interventions, and other studies comparing observational research with RCTs have used the term "concordance" instead of replicability (76). If one considers only interventional studies (i.e., RCTs and case series) in Ioannidis' study (11) – as was the case in our sample –, the replication rate is 67% when considering both significance and effect size, or 87% – exactly the same as ours when including only statistically significant highly cited studies – when considering statistical significance alone.

Defining replication boundaries

Considering a study as a replication of another inevitably requires establishing the boundary conditions of a claim (10). We opted to define replications as studies that had matching PICO general components (77). This led most highly cited studies to be classified as unchallenged, with replications being found for only 27% of our sample (as opposed to 76% in Ioannidis (11), 52% in Tajika et al. (12), and 42% in Niven et al. (13), in which criteria were less stringent). Many studies that could have been considered replications by looser criteria were thus excluded.

Even though we were more conservative than previous studies in defining replication boundaries, our study pairs were still not perfect replicas of each other. In some cases, definitions for clinical conditions were very broad, such as "cancer" in Brahmer et al. 2012 (45) and Topalian et al. 2012 (47), meaning that replication samples could potentially be quite distinct from the original one. These discrepancies thus remain as possible explanations for contradictions between results. Heterogeneity between study populations and interventions presents challenges to studying replicability in clinical research, and methodological differences between the original study and replications

seem to be common in previous studies as well (13).

Contradicted studies

Regarding our primary outcome, two phase 1 trials and two RCTs were classified as contradicted. Both phase 1 trials had CIs that did not overlap with those of the replication – in one case, the effect was larger in the original study, while in the other it was larger in the replication. As phase 1 trials typically have small sample sizes and are likely to be more prone to publication/citation bias (i.e., a negative phase 1 trial is unlikely to become highly cited), their replicability is expected to be lower than that of RCTs. Nevertheless, the majority of phase 1 trials in our sample were successfully replicated, although replication criteria were less stringent for these studies as (a) they were not subject to the statistical significance criteria and (b) CIs for their effect sizes were broader. Still, it's worth noting that all RCTs replicating phase 1 trials in our sample showed a statistically significant benefit of the intervention when compared to a control group.

Concerning RCTs, the two replication failures for initially positive studies were observed for ERSPC (29) (a large prostate cancer screening trial) and KEYNOTE-024 (68) (a trial of the checkpoint inhibitor pembrolizumab in lung cancer). ERSPC (29) was contradicted because it showed a statistically significant effect (p=0.03) while the meta-analysis did not reach significance, although effect sizes do overlap. Some methodological issues have been proposed to account for this discrepancy, such as the large degree of control group contamination in the PLCO trial (78) and the lower screening intensity in the CAP trial (79), two negative studies that account for most of the weight in the replication meta-analysis.

KEYNOTE-024 (68), meanwhile, was considered contradicted by our replication criteria, which specifically addressed the primary outcome in the highly cited study – in this case, progression-free survival. Nevertheless, both the original study and the replication (KEYNOTE-042) (69) showed an overall survival benefit, and the lack of effect on progression in the replication seems to be due to differences in the early stage of the trial, even though the intervention group fared better on the long run. Thus, the result was replicated when the more clinically relevant outcome of survival is considered, and the lack of replication for our chosen outcome appears to be a statistical

accident.

Of note, we did not use the statistical significance criterion to evaluate studies with non-significant results (as lack of statistical significance should not be taken as evidence of equivalence between treatments). Accordingly, of the two non-significant studies in our sample, one of them – the ACCORD trial (60) – had a replication with a significant result – in this case, a large meta-analysis that reached marginal significance (p=0.04). Nevertheless, other meta-analyses arrived at different conclusions (61,80–83). Effect sizes were similar between studies, suggesting that lack of agreement in this criterion was a consequence of lower statistical power in the highly cited study. PARTNER A (35), meanwhile, was a non-inferiority study that showed similar 1-year outcomes with transcatheter and surgical aortic valve replacement. Its replication (38) found a better outcome in the transcatheter group, with the authors speculating that the contrasting results could be due to the type of prosthesis used or to differences in the patients' risk profile.

Effect size inflation

Many studies analyzing replications have found evidence that published effects are systematically inflated, a fact that is expected when statistical significance thresholds are used as a criterion for publication (84). Nevertheless, strategies to measure effect size inflation vary widely across studies. Ioannidis found initially stronger effects in 7 out of 27 replicated studies, using the criteria of a decrease in risk reduction of at least 50%, or a benefit of shorter duration or limited generalizability in the replication (11). Tajika et al. reported standardized mean differences of initial studies to be 2.3 times larger than those of replications (12), while Niven et al. (13) reported a mean absolute risk difference of 16% between original studies and replications. Similar evidence of effect size inflation has also been found both in systematic replication initiatives (5,8) and meta-analyses of effect sizes over time (85,86).

Contrary to these studies, we found little evidence of systematic effect size inflation in the highly cited clinical literature between 2004 and 2018. This suggests that publication/citation bias might be more limited in our sample than in other fields of

research. That said, our ability to detect it in our primary analysis could have been reduced by the use of meta-analyses as replications, as the replication sample included the highly cited study, as well as some studies that came before it. Nevertheless, our supplementary analysis removing the primary studies from the meta-analyses found very similar inflation estimates, suggesting that this was not a major issue.

Another limitation is that using ratios for measuring effect size inflation leads to variation in estimates depending on the outcome used. If nonresponse rates or response odds were used instead of ORRs for phase 1 trials, for example, estimates of inflation increased to 17% or 18%, respectively – although CIs were still wide and compatible with absence of systematic inflation. One can also make the case that relative differences in effect sizes may be less relevant than absolute ones for clinical practice. Nevertheless, the fact that different outcome measures were used across studies makes absolute differences non-commensurable and prevents us from analyzing the sample as a whole in this manner.

Although evidence for systematic inflation was limited, this does not mean that initially stronger effects were not found in some studies, as in the case of KEYNOTE-024 (69), in which a risk reduction in progression disappeared in the replication, and TAXUS-IV (49), in which relative risk in the treated group increased from 0.39 to 0.66. Nevertheless, the fact that increases in effect size were found in other studies – such as Brahmer et al., 2012 (45), in which the ORR doubled from 13% to 27%, – suggests that some or most of these discrepancies can be explained by statistical fluctuation, and that systematic bias in favor of positive studies is smaller in this literature than in other fields.

Replication criteria

Different replication criteria complement each other by capturing distinct aspects of replicability (6,8). Statistical significance alone does not distinguish between magnitude and precision, and thus says little about how two effects compare directly (87). Comparing effect sizes, meanwhile, avoids emphasis on statistical thresholds (8), but may lead to studies with different conclusions be considered as successful replications of each other. For this reason, our primary outcome was a combination of statistical

significance and CI overlap of effect sizes.

Highly cited studies were expected to predominantly present statistically significant results for their primary outcomes, as this literature is enriched in studies with high power and high prior probabilities (88). Most replications also yielded significant results, something that would be expected if the primary findings represent true effects. In fact, considering that the replication rate for this criterion was 87% for initially positive studies, even replication failures could represent vibration around statistical thresholds (as might have been the case for the KEYNOTE-024 replication (68), for example), as studies in clinical medicine are often powered around that level.

Replicability based on CI overlap was similarly high (88%), although this is a rather loose threshold for effect size similarity: absence of CI overlap for two identical effects is expected to occur by chance alone in around 0.6% of cases (89), and this high bar for type 1 error comes at the cost of lower statistical power to detect differences in effect sizes. A more stringent criterion of having the replication effect size included in the original CI led to a lower but still reasonable (62%) replication rate, despite not considering the potential variability in replication estimates.

Inclusion of the highly cited study's point estimate in the replication CI was less frequent (38%), but this is an overly strict criterion, especially when replications have sample sizes that are much larger than the original studies.

Calculating prediction intervals for the original effect given the replication sample size would likely represent the fairest way to assess replication of effect sizes (90), but this was not always possible for all cases based on the available data.

General limitations

Defining what constitutes a replication is not trivial: even though we followed a predefined protocol to define PICO components, their abstraction inevitably involves a degree of subjectivity. Moreover, as was the case in previous studies of replications in the published literature (11–13,86), there was no way to develop a systematic search strategy that was applicable for every study. Because of these factors, our independent

searches for replications had low agreement, and a second step was needed to reach consensus. Still, it is possible that our searches could have missed some valid replication candidates.

In at least one case – the ACCORD trial on intensive glucose control (61) – there were candidate replications that reached different results (60,61,81–83). Although we used an objective criterion to define the replication selected for analysis (i.e., sample size), a replication with a different result might have been chosen if we used other criteria. Of note, we did not evaluate risk of bias or methodological quality in replications, opening up the possibility that the largest replication available might not be necessarily the most reliable one.

An important caveat in our analysis is the fact that meta-analyses were considered as replications, even though most of them included the highly cited study. This leads to a degree of circularity in the analysis that could have biased our reproducibility estimate upwards. To deal with this problem, we conducted independent meta-analyses excluding the highly cited study in order to turn them into truly independent replications. Interestingly, this did not lead to major changes in our replication rates, and actually led to lower estimates of effect size inflation. This confirms our impression that highly cited clinical literature seems to be generally replicable, and that these studies' effect sizes are not systematically higher than those of other studies on the same topic.

Even after including meta-analyses, our stringent criteria to consider studies as having matching PICO components left us with a small sample, and the high replicability rate led to an even lower number of contradicted studies. Thus, our analysis was markedly with low statistical power to detect predictors of replicability. Even though none of our predictors reached statistical significance, it seems likely that factors such as lower p values or higher sample sizes would be associated with a higher replicability rate, had a larger sample been available.

As a final limitation, we relied on replications that were published in the literature. As the existence and publication of these replications are subject both to the interest of researchers to perform them and to that of editors and reviewers to publish them, the

approach in our study is not directly comparable to the systematic replication attempts that have been performed in other areas (1–8). This is particularly important given that the majority of highly cited studies in our sample had no available replications according to our criteria – thus, selectiveness in performing or publishing replications may have biased our replicability rates. It is also possible that successfully replicated studies receive more citations in the long run, biasing the reproducibility rate of highly cited studies upwards.

Conclusions

Despite the high rate of unchallenged studies, we found the replicability rate of the highly cited clinical literature between 2004 and 2018 to be higher than previously estimated, with little evidence of effect size inflation. These numbers are valid for a narrow, very influential subsample of articles, and cannot be generalized to medical research at large. Nevertheless, they run counter to the assertion that there is a widespread reproducibility crisis in science, and suggest that this may not be the case for every scientific field.

The higher replication rate found in our study when compared to earlier samples of the clinical literature could also be taken as a sign of improvement over time; nevertheless, this conclusion is tentative at best, as differences in methodology (such as the definition of effect size inflation) and samples (such as the frequency of different study designs) do not warrant direct comparisons between studies.

Further research is warranted to examine whether the high replicability of highly cited clinical research is related to particular research practices that are not as widely used in other areas of biomedical science, such as randomization, blinding or prospective protocol registration (14,16,91–94). If such links can be reliably established, they could be used to inform attempts to improve replicability in different research fields.

Acknowledgements

An abstract for this study has been published in *BMJ Evidence-Based Medicine* (http://dx.doi.org/10.1136/bmjebm-2022-PODabstracts.105) as a product of the EBM Live 2022 Conference.

References

1. Neves K, Carneiro CF, Wasilewska-Sampaio AP, Abreu M, Valério-Gomes B, Tan PB, et al. Two years into the Brazilian Reproducibility Initiative: reflections on conducting a large-scale replication of Brazilian biomedical science. Mem Inst Oswaldo Cruz. 2020;115:e200328.

2. Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB, Alper S, et al. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. Advances in Methods and Practices in Psychological Science. 2018;1:443–90.

3. Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, Banks JB, et al. Many Labs 3: Evaluating participant pool quality across the academic semester via replication. Journal of Experimental Social Psychology. 2016;67:68–82.

4. Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, et al. Investigating the replicability of preclinical cancer biology. Pasqualini R, Franco E, editors. eLife. 2021;10:e71601.

5. Camerer CF, Dreber A, Holzmeister F, Ho TH, Huber J, Johannesson M, et al. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nat Hum Behav. 2018;2(9):637–44.

6. Cova F, Strickland B, Abatista A, Allard A, Andow J, Attie M, et al. Estimating the Reproducibility of Experimental Philosophy. Review of Philosophy and Psychology. 2021;12:9–44.

7. Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA. An open investigation of the reproducibility of cancer biology research. eLife. 2014;3.

8. Open Science Collaboration. Estimating the reproducibility of psychological science. Science. 2015 Aug 28;349(6251):aac4716.

9. Fanelli D. Opinion: Is science really facing a reproducibility crisis, and do we need it to? Proc Natl Acad Sci USA. 2018;115:2628–31.

10. Nosek BA, Errington TM. What is replication? PLOS Biology. 2020;18:e3000691.

11. Ioannidis JPA. Contradicted and Initially Stronger Effects in Highly Cited Clinical Research. JAMA. 2005;294:218–28.

12. Tajika A, Ogawa Y, Takeshima N, Hayasaka Y, Furukawa TA. Replication and contradiction of highly cited research papers in psychiatry: 10-year follow-up. The British Journal of Psychiatry. 2015;207:357–62.

13. Niven DJ, McCormick TJ, Straus SE, Hemmelgarn BR, Jeffs L, Barnes TRM, et al. Reproducibility of clinical research in critical care: a scoping review. BMC Medicine. 2018;16:26.

14. Zarin DA, Tse T, Williams RJ, Rajakannan T. Update on Trial Registration 11 Years after the ICMJE Policy Was Established. New England Journal of Medicine. 2017;376:383–91.

15. Goldacre B, Drysdale H, Dale A, Milosevic I, Slade E, Hartley P, et al. COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. Trials. 2019;20:118.

16. Kaplan RM, Irvin VL. Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. PLOS ONE. 2015;10:e0132382.

17. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D, et al. The Revised CONSORT Statement for Reporting Randomized Trials: Explanation and Elaboration. Ann Intern Med. 2001;134:663–94.

18. Schulz KF, Altman DG, Moher D, the CONSORT Group. CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMC Medicine. 2010;8:18.

19. Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A, et al. Data Sharing Statements for Clinical Trials — A Requirement of the International Committee of Medical Journal Editors. N Engl J Med. 2017;376:2277–9.

20. Kim C, Prasad V. Cancer Drugs Approved on the Basis of a Surrogate End Point and Subsequent Overall Survival: An Analysis of 5 Years of US Food and Drug Administration Approvals. JAMA Internal Medicine. 2015;175:1992–4.

21. Ebrahim S, Sohani ZN, Montoya L, Agarwal A, Thorlund K, Mills EJ, et al. Reanalyses of Randomized Clinical Trial Data. JAMA. 2014;312:1024–32.

22. Jager LR, Leek JT. An estimate of the science-wise false discovery rate and application to the top medical literature. Biostatistics. 2014;15:1–12.

23. R Core Team C. R Software [Internet]. Vienna, Austria; (R: A language and environment for statistical computing.). Available from: https://www.R-project.org/

24. Chapman PB, Hauschild A, Robert C, Haanen JB, Ascierto P, Larkin J, et al. Improved Survival with Vemurafenib in Melanoma with BRAF V600E Mutation. N Engl J Med. 2011 Jun 30;364(26):2507–16.

25. Larkin J, Chiarion-Sileni V, Gonzalez R, Grob JJ, Cowey CL, Lao CD, et al. Combined Nivolumab and Ipilimumab or Monotherapy in Untreated Melanoma. N Engl J Med. 2015;373:23–34.

26. Campbell BCV, Mitchell PJ, Kleinig TJ, Dewey HM, Churilov L, Yassi N, et al. Endovascular Therapy for Ischemic Stroke with Perfusion-Imaging Selection. N Engl J Med. 2015;372:1009–18.

27. Lieberman JA, Stroup TS, McEvoy JP, Swartz MS, Rosenheck RA, Perkins DO, et al. Effectiveness of Antipsychotic Drugs in Patients with Chronic Schizophrenia. N Engl J Med. 2005;353:1209–23.

28. Cochrane. Cochrane Handbook for Systematic Reviews of Interventions | Cochrane Training [Internet]. Available from: https://training.cochrane.org/handbook/current

29. Schröder FH, Hugosson J, Roobol MJ, Tammela TLJ, Ciatto S, Nelen V, et al. Screening and Prostate-Cancer Mortality in a Randomized European Study. N Engl J Med. 2009;360:1320–8.

30. Rosell R, Carcereny E, Gervais R, Vergnenegre A, Massuti B, Felip E, et al. Erlotinib versus standard chemotherapy as first-line treatment for European patients with advanced EGFR mutation-positive non-small-cell lung cancer (EURTAC): a multicentre, open-label, randomised phase 3 trial. The Lancet Oncology. 2012;13:239–46.

31. Mok TS, Wu YL, Thongprasert S, Yang CH, Chu DT, Saijo N, et al. Gefitinib or Carboplatin–Paclitaxel in Pulmonary Adenocarcinoma. N Engl J Med. 2009;361:947–57.

32. Maemondo M, Inoue A, Kobayashi K, Sugawara S, Oizumi S, Isobe H, et al. Gefitinib or Chemotherapy for Non–Small-Cell Lung Cancer with Mutated EGFR. N Engl J Med. 2010;362:2380–8.

33. Cheng AL, Kang YK, Chen Z, Tsao CJ, Qin S, Kim JS, et al. Efficacy and safety of sorafenib in patients in the Asia-Pacific region with advanced hepatocellular carcinoma: a phase III randomised, double-blind, placebo-controlled trial. The Lancet Oncology. 2009;10:25–34.

34. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. BMJ Evidence-Based Medicine. 2016;21:125–7.

35. Smith CR, Leon MB, Mack MJ, Miller DC, Moses JW, Svensson LG, et al. Transcatheter versus Surgical Aortic-Valve Replacement in High-Risk Patients. N Engl J Med. 2011;364:2187–98.

36. Pagnesi M, Chiarito M, Stefanini GG, Testa L, Reimers B, Colombo A, et al. Is Transcatheter Aortic Valve Replacement Superior to Surgical Aortic Valve Replacement?: A Meta-Analysis of Randomized Controlled Trials. JACC: Cardiovascular Interventions. 2017;10:1899–901.

37. Greenhalgh J, Boland A, Bates V, Vecchio F, Dundar Y, Chaplin M, et al. First-line treatment of advanced epidermal growth factor receptor (EGFR) mutation positive non-squamous non-small cell lung cancer. Cochrane Database of Systematic Reviews. 2021;(3).

38. Adams DH, Popma JJ, Reardon MJ, Yakubov SJ, Coselli JS, Deeb GM, et al. Transcatheter Aortic-Valve Replacement with a Self-Expanding Prosthesis. N Engl J Med. 2014;370:1790–8.

39. Schwarzer G. meta: General Package for Meta-Analysis [Internet]. 2023. Available from: https://CRAN.R-project.org/package=meta

40. Higgins JPT, Thompson SG, Spiegelhalter DJ. A Re-Evaluation of Random-Effects Meta-Analysis. Journal of the Royal Statistical Society Series A (Statistics in Society). 2009;172:137–59.

41. Hamid O, Robert C, Daud A, Hodi FS, Hwu WJ, Kefford R, et al. Safety and Tumor Responses with Lambrolizumab (Anti–PD-1) in Melanoma. N Engl J Med. 2013;369:134–44.

42. Niu M, Hong D, Ma TC, Chen XW, Han JH, Sun J, et al. Short-term and long-term efficacy of 7 targeted therapies for the treatment of advanced hepatocellular carcinoma: a network meta-analysis: Efficacy of 7 targeted therapies for AHCC. Medicine. 2016;95:e5591.

43. Rodrigues FB, Neves JB, Caldeira D, Ferro JM, Ferreira JJ, Costa J. Endovascular treatment versus medical care alone for ischaemic stroke: systematic review and meta-analysis. BMJ. 2016;353:i1754.

44. Zhao Y, Liu J, Cai X, Pan Z, Liu J, Yin W, et al. Efficacy and safety of first line treatments for patients with advanced epidermal growth factor receptor mutated, non-small cell lung cancer: systematic review and network meta-analysis. BMJ. 2019;367:l5460.

45. Brahmer JR, Tykodi SS, Chow LQM, Hwu WJ, Topalian SL, Hwu P, et al. Safety and Activity of Anti–PD-L1 Antibody in Patients with Advanced Cancer. N Engl J Med. 2012;366:2455–65.

46. Zhang T, Xie J, Arai S, Wang L, Shi X, Shi N, et al. The efficacy and safety of anti-PD-1/PD-L1 antibodies for treatment of advanced or refractory cancers: a meta-analysis. Oncotarget. 2016;(45):73068–79.

47. Topalian SL, Hodi FS, Brahmer JR, Gettinger SN, Smith DC, McDermott DF, et al. Safety, Activity, and Immune Correlates of Anti–PD-1 Antibody in Cancer. N Engl J Med. 2012;366:2443–54.

48. Tie Y, Ma X, Zhu C, Mao Y, Shen K, Wei X, et al. Safety and efficacy of nivolumab in the treatment of cancers: A meta-analysis of 27 prospective clinical trials. International Journal of Cancer. 2017;140:948–58.

49. Stone GW, Ellis SG, Cox DA, Hermiller J, O'Shaughnessy C, Mann JT, et al. A Polymer-Based, Paclitaxel-Eluting Stent in Patients with Coronary Artery Disease. N Engl J Med. 2004;350:221–31.

50. Bangalore S, Toklu B, Amoroso N, Fusaro M, Kumar S, Hannan EL, et al. Bare metal stents, durable polymer drug eluting stents, and biodegradable polymer drug eluting stents for coronary artery disease: mixed treatment comparison meta-analysis. BMJ. 2013;347:f6625.

51. Serruys PW, Morice MC, Kappetein AP, Colombo A, Holmes DR, Mack MJ, et al. Percutaneous Coronary Intervention versus Coronary-Artery Bypass Grafting for Severe Coronary Artery Disease. N Engl J Med. 2009;360:961–72.

52. Ali WE, Vaidya SR, Ejeh SU, Okoroafor KU. Meta-analysis study comparing percutaneous coronary intervention/drug eluting stent versus coronary artery bypass surgery of unprotected left main coronary artery disease: Clinical outcomes during short-term versus long-term (> 1 year) follow-up. Medicine. 2018;97:e9909.

53. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A, Untch M, Smith I, et al. Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer. N Engl J Med. 2005;353:1659–72.

54. Genuino AJ, Chaikledkaew U, The DO, Reungwetwattana T, Thakkinstian A. Adjuvant trastuzumab regimen for HER2-positive early-stage breast cancer: a systematic review and meta-analysis. Expert Review of Clinical Pharmacology. 2019;12:815–24.

55. Soares-Weiser K, Béchard-Evans L, Howard Lawson A, Davis J, Ascher-Svanum H. Time to all-cause treatment discontinuation of olanzapine compared to other antipsychotics in the treatment of schizophrenia: A systematic review and meta-analysis. European Neuropsychopharmacology. 2013;23:118–25.

56. Llovet JM, Ricci S, Mazzaferro V, Hilgard P, Gane E, Blanc JF, et al. Sorafenib in Advanced Hepatocellular Carcinoma. N Engl J Med. 2008;359:378–90.

57. Ilic D, Djulbegovic M, Jung JH, Hwang EC, Zhou Q, Cleves A, et al. Prostate cancer screening with prostate-specific antigen (PSA) test: a systematic review and meta-analysis. BMJ. 2018;362:k3519.

58. Shaw AT, Kim DW, Nakagawa K, Seto T, Crinó L, Ahn MJ, et al. Crizotinib versus Chemotherapy in Advanced ALK-Positive Lung Cancer. N Engl J Med. 2013;368:2385–94.

59. Elliott J, Bai Z, Hsieh SC, Kelly SE, Chen L, Skidmore B, et al. ALK inhibitors for non-small cell lung cancer: A systematic review and network meta-analysis. PLOS ONE. 2020;15:e0229179.

60. The Action to Control Cardiovascular Risk in Diabetes Study Group TA to CCR in DSG, Gerstein H, Miller M, Byington R, Goff Jr D, Bigger T, et al. Effects of Intensive Glucose Lowering in Type 2 Diabetes. N Engl J Med. 2008;358:2545–59.

61. Fang HJ, Zhou YH, Tian YJ, Du HY, Sun YX, Zhong LY. Effects of intensive glucose lowering in treatment of type 2 diabetes mellitus on cardiovascular outcomes: A meta-analysis of data from 58,160 patients in 13 randomized controlled trials. International Journal of Cardiology. 2016;218:50–8.

62. Goyal M, Menon BK, Zwam WH van, Dippel DWJ, Mitchell PJ, Demchuk AM, et al. Endovascular thrombectomy after large-vessel ischaemic stroke: a meta-analysis of individual patient data from five randomised trials. The Lancet. 2016;387:1723–31.

63. Berkhemer OA, Fransen PSS, Beumer D, van den Berg LA, Lingsma HF, Yoo AJ, et al. A Randomized Trial of Intraarterial Treatment for Acute Ischemic Stroke. N Engl J Med. 2015;372:11–20.

64. Hacke W, Kaste M, Bluhmki E, Brozman M, Dávalos A, Guidetti D, et al. Thrombolysis with Alteplase 3 to 4.5 Hours after Acute Ischemic Stroke. N Engl J Med. 2008;359:1317–29.

65. Wardlaw JM, Murray V, Berge E, Zoppo G del, Sandercock P, Lindley RL, et al. Recombinant tissue plasminogen activator for acute ischaemic stroke: an updated systematic review and meta-analysis. The Lancet. 2012;379:2364–72.

66. Goyal M, Demchuk AM, Menon BK, Eesa M, Rempel JL, Thornton J, et al. Randomized Assessment of Rapid Endovascular Treatment of Ischemic Stroke. N Engl J Med. 2015;372:1019–30.

67. Pyo JS, Kang G. Immunotherapy in advanced melanoma: a network meta-analysis. Immunotherapy. 2017;9:471–9.

68. Reck M, Rodríguez-Abreu D, Robinson AG, Hui R, Csőszi T, Fülöp A, et al. Pembrolizumab versus Chemotherapy for PD-L1–Positive Non–Small-Cell Lung Cancer. N Engl J Med. 2016;375:1823–33.

69. Mok TSK, Wu YL, Kudaba I, Kowalski DM, Cho BC, Turna HZ, et al. Pembrolizumab versus chemotherapy for previously untreated, PD-L1-expressing, locally advanced or metastatic non-small-cell lung cancer (KEYNOTE-042): a randomised, open-label, controlled, phase 3 trial. The Lancet. 2019;393:1819–30.

70. Garon EB, Rizvi NA, Hui R, Leighl N, Balmanoukian AS, Eder JP, et al. Pembrolizumab for the Treatment of Non–Small-Cell Lung Cancer. N Engl J Med. 2015;372:2018–28.

71. Herbst RS, Baas P, Kim DW, Felip E, Pérez-Gracia JL, Han JY, et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. The Lancet. 2016;387:1540–50.

72. Wolchok JD, Kluger H, Callahan MK, Postow MA, Rizvi NA, Lesokhin AM, et al. Nivolumab plus Ipilimumab in Advanced Melanoma. N Engl J Med. 2013;369:122–33.

73. Flaherty KT, Puzanov I, Kim KB, Ribas A, McArthur GA, Sosman JA, et al. Inhibition of Mutated, Activated BRAF in Metastatic Melanoma. N Engl J Med. 2010;363:809–19.

74. Fong PC, Boss DS, Yap TA, Tutt A, Wu P, Mergui-Roelvink M, et al. Inhibition of Poly(ADP-Ribose) Polymerase in Tumors from BRCA Mutation Carriers. N Engl J Med. 2009;361:123–34.

75. Kaufman B, Shapira-Frommer R, Schmutzler RK, Audeh MW, Friedlander M, Balmaña J, et al. Olaparib Monotherapy in Patients With Advanced Cancer and a Germline BRCA1/2 Mutation. Journal of Clinical Oncology. 2015;33:244–50.

76. Moorthy D, Chung M, Lee J, Yu WW, Lau J, Trikalinos TA. Concordance Between the Findings of Epidemiological Studies and Randomized Trials in Nutrition: An Empirical Evaluation and Citation Analysis. Agency for Healthcare Research and Quality (US); 2013.

77. Schwingshackl L, Balduzzi S, Beyerbach J, Bröckelmann N, Werner SS, Zähringer J, et al. Evaluating agreement between bodies of evidence from randomised controlled trials and cohort studies in nutrition research: meta-epidemiological study. BMJ. 2021;374:n1864.

78. Tsodikov A, Gulati R, Heijnsdijk EAM, Pinsky PF, Moss SM, Qiu S, et al. Reconciling the Effects of Screening on Prostate Cancer Mortality in the ERSPC and PLCO Trials. Ann Intern Med. 2017;167(7):449–55.

79. Martin RM, Donovan JL, Turner EL, Metcalfe C, Young GJ, Walsh EI, et al. Effect of a Low-Intensity PSA-Based Screening Intervention on Prostate Cancer Mortality: The CAP Randomized Clinical Trial. JAMA. 2018;319:883–95.

80. Anushka P, S M, J C, B N, L B, M W, et al. Intensive blood glucose control and vascular outcomes in patients with type 2 diabetes. N Engl J Med. 2008;358(24).

81. Hemmingsen B, Lund SS, Gluud C, Vaag A, Almdal T, Hemmingsen C, et al. Intensive glycaemic control for patients with type 2 diabetes: systematic review with meta-analysis and trial sequential analysis of randomised clinical trials. BMJ. 2011;343:d6898.

82. Marso SP, Kennedy KF, House JA, McGuire DK. The effect of intensive glucose control on all-cause and cardiovascular mortality, myocardial infarction and stroke in persons with type 2 diabetes mellitus: a systematic review and meta-analysis. Diabetes and Vascular Disease Research. 2010;7:119–30.

83. Sardar P, Udell JA, Chatterjee S, Bansilal S, Mukherjee D, Farkouh ME. Effect of Intensive Versus Standard Blood Glucose Control in Patients With Type 2 Diabetes Mellitus in Different Regions of the World: Systematic Review and Meta-analysis of Randomized Controlled Trials. Journal of the American Heart Association. 2015;4:e001577.

84. Ioannidis JPA. Why Most Discovered True Associations Are Inflated. Epidemiology. 2008;19:640–8.

85. Fanelli D, Costas R, Ioannidis JPA. Meta-assessment of bias in science. PNAS. 2017;114:3714–9.

86. Cristea IA, Georgescu R, Ioannidis JPA. Effect Sizes Reported in Highly Cited Emotion Research Compared With Larger Studies and Meta-Analyses Addressing the Same Questions. Clinical Psychological Science. 2022;10:786–800.

87. Amrhein V, Korner-Nievergelt F, Roth T. The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. PeerJ. 2017;5:e3544.

88. Ioannidis JPA. Why Most Published Research Findings Are False. PLOS Medicine. 2005;2:e124.

89. Austin PC, Hux JE. A brief note on overlapping confidence intervals. Journal of Vascular Surgery. 2002;36:194–5.

90. Spence JR, Stanley DJ. Prediction Interval: What to Expect When You're Expecting … A Replication. PLOS ONE. 2016;11:e0162874.

91. Chan AW, Altman DG. Epidemiology and reporting of randomised trials published in PubMed journals. The Lancet. 2005;(9465):1159–62.

92. Protzko J, Krosnick J, Nelson LD, Nosek BA, Axt J, Berent M, et al. High Replicability of Newly-Discovered Social-behavioral Findings is Achievable [Internet]. PsyArXiv; 2020. Available from: 10.31234/osf.io/n2a9x

93. Serghiou S, Contopoulos-Ioannidis DG, Boyack KW, Riedel N, Wallach JD, Ioannidis JPA. Assessment of transparency indicators across the biomedical literature: How open is open? PLOS Biology. 2021;19:e3001107.

94. Macleod MR, McLean AL, Kyriakopoulou A, Serghiou S, Wilde A de, Sherratt N, et al. Risk of Bias in Reports of In Vivo Research: A Focus for Improvement. PLOS Biology. 2015;13:e1002273.