Strength of Statistical Evidence for the Efficacy of Cancer Drugs: A Bayesian Re-Analysis of Trials Supporting FDA Approval.

Merle-Marie Pittelkow[1,2 *] (0000-0002-7487-7898), Maximilian Linde[1] (0000-0001-8421-090X), Ymkje Anna de Vries[3] (0000-0003-4580-4873), Lars G. Hemkens[4,5,6,7] (0000-0002-3444-1432), Andreas M. Schmitt[5,8] (0000-0002-9568-8164), Rob R. Meijer[1] (0000-0001-5368-992X), & Don van Ravenzwaaij[1] (0000-0002-5030-4091)

[1] Faculty of Behavioural and Social Sciences, University of Groningen, Groningen, the Netherlands

[2] QUEST Center for Responsible Research, Berlin Institute of Health at Charité, Berlin, Germany

[3] Department of Child and Adolescent Psychiatry, University of Groningen, University Medical Center Groningen, Groningen, The Netherlands

[4] Research Center for Clinical Neuroimmunology and Neuroscience Basel (RC2NB), University Hospital Basel and University of Basel, Basel, Switzerland

[5] Department of Clinical Research, University of Basel and University Hospital Basel, Basel, Switzerland

[6] Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Stanford, California, USA

[7] Meta-Research Innovation Center Berlin (METRIC-B), Berlin Institute of Health, Berlin, Germany

[8] Department of Medical Oncology, University Hospital Basel, Basel, Switzerland.

*Correspondence to: merle-marie.pittelkow@bih-charite.de , Anna-Louisa-Karsch Straße 2, 10178 Berlin, Germany

ABSTRACT

Objective: To quantify the strength of statistical evidence of randomised controlled trials (RCTs) for novel cancer drugs approved by the Food and Drug Administration (FDA) in the last two decades.

Methods: We used data on overall survival (OS), progression-free survival (PFS), and tumour response (TR) for novel cancer drugs approved for the first time by the FDA between January 2000 and December 2020. We assessed strength of statistical evidence by calculating Bayes Factors (*BF*s) for all available endpoints, and we pooled evidence using Bayesian fixed-effect meta-analysis for indications approved based on two RCTs. Strength of statistical evidence was compared between endpoints, approval pathways, lines of treatment, and types of cancer.

Results: We analysed the available data from 82 RCTs corresponding to 68 indications supported by a single RCT and seven indications supported by two RCTs. Median strength of statistical evidence was ambiguous for OS (*BF* = 1.9; IQR 0.5-14.5), and strong for PFS (*BF* = 24,767.8; IQR 109.0-7.3*10^6) and TR (*BF* = 113.9; IQR 3.0-547,100). Overall, 44 indications (58.7%) were approved without clear statistical evidence for OS improvements and seven indications (9.3%) were approved without statistical evidence for improvements on any endpoint. Strength of statistical evidence was lower for accelerated approval compared to non-accelerated approval across all three endpoints. No meaningful differences were observed for line of treatment and cancer type.

Limitations: This analysis is limited to statistical evidence. We did not consider non-statistical factors (e.g., risk of bias, quality of the evidence).

Discussion: *BF*s offer novel insights into the strength of statistical evidence underlying cancer drug approvals. Most novel cancer drugs lack strong statistical evidence that they improve OS, and a few lack statistical evidence for efficacy altogether. These cases require a transparent and clear explanation. When evidence is ambiguous, additional post-marketing trials could eliminate uncertainty.

SUMMARY BOX

What is already known about the topic.

Novel cancer drugs are approved based on limited evidence.

Between 2000 and 2020, novel cancer drugs were typically approved based on a single pivotal trial, with only half of these being randomised controlled trials and many relying on surrogate endpoints to demonstrate efficacy.

What this study adds.

Using a newly developed Bayesian survival analysis strategy, we provide novel insights into the strength of statistical evidence given the limited data available at approval. Most novel cancer drugs lack strong statistical evidence that they improve overall survival, and a minority lack evidence for efficacy altogether.

BACKGROUND

For a new cancer drug to be marketed in the U.S., it needs to be endorsed by the Food and Drug Administration (FDA). Cancer drugs are frequently approved based on limited evidence, which increases uncertainty in clinical decision making [1–6]. Despite FDA guidelines suggesting that substantial evidence for efficacy based on two convincing trials should be provided [7], approval for novel cancer drugs between 2000 and 2020 was typically based on a single pivotal (i.e., efficacy determining) trial [1,6]. The understanding that this leads to a reduced level of evidential strength is implicit, but explicit quantification for the available strength of evidence is missing. A further complication is the fact that the FDA allows the use of surrogate endpoints, that is, outcomes that are thought to be associated with clinically relevant outcomes such as overall survival (OS) but are easier or faster to measure [8]. Examples of such surrogate endpoints, like progression-free survival (PFS) and tumour response (TR), may not be true predictors of the clinical outcome of interest [8–10], limiting the quality of the evidence available at the time of approval. It has been questioned whether the current state of evidence for cancer medication fulfils the FDA's requirement of demonstrating "meaningful therapeutic benefits" [2].

To allow for informed regulatory decision making, efficacy determinations should reflect the evidence at the time of approval. Bayes factors (*BF*s) provide a relative measure of statistical evidence for or against competing hypotheses. Technically, it is comparing the likelihood of the observed data under the null hypothesis (i.e., no treatment effect) to the marginal likelihood of the observed data under an alternative hypothesis (e.g., a positive treatment effect; one-sided alternative hypothesis) [11–13]. The resulting ratio provides a continuous measure of the strength of statistical evidence (see Figure 1). For example, *BF*s can be used to compare the hypothesis that a novel treatment improves OS (often referred to as the alternative hypothesis) and the hypothesis that the novel treatment does not improve OS (often referred to as the null hypothesis). A resulting *BF* of 1 would indicate complete ambiguity as the observed study results are equally likely to have occurred if the novel treatment improves OS or if the novel treatment does not improve OS. A resulting *BF* of 10 would indicate that the observed study results are ten times more likely to have occurred given that the novel treatment improves OS than if it does not improve OS, whereas a resulting *BF* of 0.1 would indicate that the observed study results are ten times more likely to have occurred given that the novel treatment does not work than if it does work.

[FIGURE 1 HERE]

Using a Bayesian framework we offer a different perspective on approval evidence that may be more intuitive than the traditional focus on statistical significance in frequentist frameworks. We take a perspective similar to that of a clinician who wants to know if the results of a diagnostic test

(here, the approval trials) are more likely under a working diagnosis than under a differential diagnosis, which is reflected by the *BF*. This project has two aims. First, we aim to describe the strength of statistical evidence associated with novel cancer drugs approved for the first time between 2000 and 2020 using a Bayesian framework. Different standards of how much uncertainty is acceptable and what constitutes "sufficient" strength of statistical evidence for beneficial effects (and against harmful effects) at the time of approval might exist. For example, one might accept more uncertainty and lower strength of statistical evidence if the indication concerns a 3rd or 4th line treatment instead of 1st or 2nd. Moreover, strength of statistical evidence might also vary between cancer types. Previous reports suggest that novel drugs for haematological cancers are more likely to be approved based on single arm trials and surrogate endpoints compared to solid cancers [1]. Therefore, our second aim is to contrast strength of statistical evidence between endpoints (OS vs. PFS vs. TR), approval types (accelerated vs. non-accelerated), lines of treatment (1st vs. 2nd vs. 3rd and up), and type of cancer (solid vs. hematologic) to describe potential differences in strength of statistical evidence.

METHOD

Ethics Approval

This study involved publicly available trial-level data. No ethical approval was needed.

Data and Registration

The aim and general approach of the project were registered at OSF (https://osf.io/exyfd/). Analysis code and data are publicly available from https://ceit-cancer.org/ and the OSF framework (https://osf.io/4uhz7). All files needed to reproduce the reported analyses are available on the OSF framework (https://osf.io/qz7xy/).

FDA Data

We used data from the CEIT-Cancer project (details provided elsewhere [14]). In short, novel drugs and biological therapies receiving first approval for the treatment of any malignant diseases between January 2000 and December 2020 and the corresponding FDA reviews (available at drugs@FDA) were identified. For all RCTs evaluating the drug in the approved indication (regardless of whether the trial was described as pivotal or not) and for all single-arm clinical trials that were explicitly described as pivotal, the following data were extracted: hazard ratios (HR) and associated 95% confidence intervals for OS and/or PFS; median OS and/or PFS; number of events for OS, PFS and/or TR; sample size; line of treatment; approval pathway (i.e., priority review, orphan design, accelerated approval, and breakthrough therapy designation); type of cancer; type of control; and type of blinding.

Data Analysis

Data analysis was performed in R version 4.1.2 [15] . We calculated *BF*s for all available endpoints per trial. For OS and PFS, we used the available summary statistics (i.e, HR, confidence interval of HR, sample size, and number of events per group) to conduct a Bayesian Cox regression using the "baymedr" R package [16]. HRs and their corresponding confidence intervals were log-transformed to obtain b coefficients and their corresponding confidence intervals. We used a standard normal distribution that was truncated at 0 (i.e., densities of 0 for values higher than 0) as the prior for beta under the alternative hypothesis. To conduct a Bayesian Cox regression without access to the full data set, multiple survival datasets were simulated based on the HR, associated confidence intervals, sample size, and number of events in each group (see [17] for details). For each simulated dataset, one *BF* was calculated. Our preregistration did not specify the number of simulated datasets to be used. We compared the results from analyses based on 45 simulated datasets to those based on 100 simulated datasets (see Figure S1 and S2 in the supplement) and found that qualitative interpretation of results did not depend on the number of simulations; we here present results only for analyses based on 100 simulated datasets. For TR outcomes, *BF*s for chi-square tests were calculated using the "BayesFactor" R package [18]. Under the null hypothesis, the prior for the proportion of TR was a joint uniform distribution ranging from 0 to 1. Under the alternative hypothesis, the prior for the proportion of TR was a uniform distribution ranging from 0 to 1 for each group independently [19].

For approvals based on two RCTs**,** we combined the available data from both trials to obtain an overall measure of the strength of statistical evidence underpinning the approval decisions. To this end, we pooled available outcomes for OS and PFS via fixed-effect meta-analysis using the "metaBMA" R package [20]. Under the alternative hypothesis, the prior for the treatment effect was a default Cauchy distribution that was truncated at 0 (i.e., densities of 0 for values higher than 0), with a location parameter of 0 and a scale parameter of $\frac{1}{\sqrt{2}}$ [21]. For TR, we pooled the number of events and number of participants and calculated corresponding chi-square tests using the "BayesFactor" R package with the same specifications as for single trial results.

The calculated *BF*s quantify the likelihood of the data under the alternative hypothesis that the treatment group performs better than the control group relative to the likelihood of the data under the null hypothesis that the treatment and control group perform equally. In other words, we expected a higher hazard (for OS and PFS) or higher event count (for TR) in the control as compared to the treatment group. To describe the strength of statistical evidence we adopted the following thresholds: *BF* below ⅓ are defined as *pro-null* evidence (i.e., evidence that the control group patients fare better or not worse), *BF* between ⅓ and 3 are defined as *ambiguous* evidence, *BF*

between 3 and 10 are defined as *moderate pro-alternative* evidence (i.e., moderate evidence that the patients with the novel drug treatment fare better), *BF* between 10 and 30 are defined as *strong pro-alternative* evidence, and *BF* above 30 are defined as *very strong pro-alternative* evidence [13]. These thresholds have been proposed by methodological researchers and are not clinically informed [13,22].

The interpretation of pro-null *BF*s depends on the nature of the control group. In placebo-controlled trials, a pro-null *BF* means that the drug likely performs comparable to or worse than placebo, indicating no efficacy of the novel drug. In trials with an active control group (e.g., an established treatment), interpreting pro-null *BF*s is more difficult. Assuming that the active control group receives an effective therapy, a pro-null *BF* does not necessarily mean that the treatment with the novel drug is ineffective, but only that it is not more effective than the control group treatment. In the remainder of this article, we therefore discuss evidence per control condition whenever possible. For the subgroup analysis of approval pathway, line of treatment, and type of cancer, we did not separate results by control group as these subgroups were too small to be split further. This decision was not pre-registered.

Exploratory data analysis

We conducted three not pre-registered, exploratory analyses. First, *BF*s offer additional information beyond effect size. Therefore, we examined the association between *BF, HR,* and the associated confidence intervals visually and via counts. To this end, we differentiated between effect sizes for which the confidence interval excluded one and those that included one, as they indicate uncertainty regarding the effect size. If confidence intervals were not available (n=2), we assumed that the confidence interval excluded 1. Second, all else being equal, greater sample sizes cause the *BF* to tend to become more extreme (i.e., tend toward 0 if the null hypothesis is supported or toward infinity if the alternative hypothesis is supported). For example, if the data is more in line with the alternative hypothesis (i.e., the novel treatment works) more data is going to push the *BF* more towards infinity. However, this relationship is not straightforward and ambiguous evidence is not always due to small sample sizes. Therefore, we examined the relationship between *BF*s and sample size both visually and with a correlation. Third, we conducted an exploratory, qualitative analysis to understand the reasoning behind the seven endorsement decisions approved based on pro-null or ambiguous statistical evidence. The methods for the qualitative analysis and a detailed account for each individual drug are discussed in the supplement (section 3).

RESULTS

The initial dataset contained data on 145 novel cancer drugs for 156 indications based on 186 trials. Almost half of the indications (n = 75; 48.1%) received FDA approval without supporting

evidence from at least one RCT. The remaining 81 indications were primarily supported by a single RCT (*n* = 70; 44.9%), with 11 (7.1%) indications supported by two RCTs. For 10 RCTs, data was unavailable.

Consequently, we analysed 82 RCTs supporting approval for 75 indications of 75 novel cancer drugs, of which 68 (90.7%) were supported by a single RCT, and 7 (8.5%) were supported by two RCTs. Of the 82 included trials, 72 trials assessed OS (87.8%), 66 trials assessed PFS (80.5%), and 68 trials assessed TR (82.9%). Overall, 52 trials provided data for all three endpoints (63.4%), 20 trials data for two endpoints (24.4%), and 10 trials for only one endpoint (12.2%; Table 1). Trial characteristics and individual *BF*s per trial and endpoint are presented in the supplement (Table S1 and S2).

[TABLE 1 HERE]

Approval decisions based on one RCT

Endpoints

Statistical evidence for a beneficial effect on at least one endpoint was moderate for five out of 68 (7.4%), and (very) strong for 49 out of 68 (86.8%) indications. The median *BF* for OS was 1.9 (IQR 0.5-14.5), for PFS 24,767.8 (IQR 109.0-7.3*10^6) and TR 113.9 (IQR 3.0 - 547,100) (Figure 2).

[FIGURE 2 HERE]

Active control groups.

Out of the 68 indications, 22 (32.4%) were approved based on trials with active control groups (e.g., established treatments) with a median *BF* of 17.9 (IQR 1.0 - 12,220.0) across all three endpoints. Of these, three (13.6%) indications lacked strong or very strong statistical evidence for a beneficial effect on any endpoint, while five indications (22.7%) were approved based on strong or very strong statistical evidence for benefits across all three endpoints.

Of the 22 indications, 19 (90.9%) considered OS, 19 (90.9%) considered PFS, and 21 (95.5%) considered TR. Of 19 indications with data available for OS, two (10.5%) were approved with pro-null or ambiguous statistical evidence for OS improvements without (very) strong evidence for improvements on surrogate endpoints. An additional 12 drugs (63.2%) were approved based on pro-null or ambiguous statistical evidence for OS improvements but with (very) strong statistical evidence for improvements on at least one surrogate outcome. One indication (5.3%) was approved based on moderate statistical evidence for OS improvements and (very) strong evidence for improvements on the surrogate outcomes. The remaining four (21.1%) indications were approved based on (very) strong statistical evidence for OS improvements.

For PFS or TR outcomes, the majority of trials indicated (very) strong statistical evidence for better treatment outcomes in the experimental compared to the control group (*n* = 15, 79.0% and *n*

231 = 16, 76.2% respectively; see also Figure 3). One drug (i.e., dasatinib) was approved based on
232 ambiguous statistical evidence for TR only.

233 **[FIGURE 3 HERE]**

234 ***Inactive (placebo) control groups.***

235 An additional 46 (67.7%) indications were approved based on trials with inactive control
236 groups (29 trials [42.7%] compared the treatment to a placebo control group; 17 trials [25.0%]
237 compared the treatment plus usual/supportive care to usual/supportive care alone) with a median
238 *BF* of 58.2 (IQR 2 - 47,080.0) across all three endpoints. Of these, two (4.3%) indications lacked
239 strong evidence for a beneficial effect on any endpoint (thus questioning the efficacy of the drug),
240 while seven (17.1%) were approved based on strong or very strong statistical evidence for benefits
241 across all three endpoints.

242 Of the 46 indications approved based on inactive-comparator trials, 41 (89.1%) considered
243 OS, 38 (84.8%) considered PFS, and 36 (78.26%) considered TR. Of 41 indications with data available
244 for OS, two (4.9%) were approved with pro-null or ambiguous statistical evidence for improvements
245 of OS and all other endpoints. An additional 18 drugs (43.9%) were approved based on pro-null or
246 ambiguous statistical evidence for OS improvements but with (very) strong statistical evidence for
247 improvements on at least one surrogate outcome. Eight (19.5%) indications were approved based on
248 moderate statistical evidence for OS improvements, of which five were supplemented with very
249 strong statistical evidence for improvements of at least one surrogate endpoint. The remaining 13
250 (31.7%) indications were supported by (very) strong statistical evidence for OS.

251 The majority of indications were supported by very strong statistical evidence for
252 improvements of PFS or TR (*n* = 34, 89.5% and *n* = 24, 66.7% respectively; see also Figure 3). One
253 drug (neratinib maleate) was approved based on moderate statistical evidence for PFS
254 improvements.

255 **Accelerated approval**

256 Data was available for seven accelerated approval decisions (Figure 4). Strength of statistical
257 evidence was consistently lower for accelerated approvals across endpoints. Overall, there were no
258 indications with accelerated approval that either lacked or provided strong or very strong statistical
259 evidence for improvements across all three endpoints. For indications that received non-accelerated
260 approval, 5 of 58 indications lacked strong evidence for improvements on any endpoint, while 12
261 (out of 58) had strong or very strong statistical evidence for improvements across all three
262 endpoints.

263 Five out of seven (71.4%) accelerated approval decisions were based on pro-null or
264 ambiguous statistical evidence for OS improvements, compared to 29 out of 53 (54.7%) non-

265 accelerated approval decisions. Two (20.0%) indications receiving accelerated approval were based
266 on pro-null evidence for OS improvements. For one drug (i.e., nivolumab) the treatment did not
267 perform better than the active comparator, whereas for the other drug (i.e., panitumumab) the
268 treatment did not outperform a placebo control.

269 For PFS, no trial provided pro-null statistical evidence. However, the proportion of
270 ambiguous statistical evidence was higher for accelerated approval (28.6 %) compared to non-
271 accelerated approval (10.0%). For TR, the proportion of pro-null evidence was similar for both
272 accelerated (14.3%) and non-accelerated approval (14.0%), but accelerated approval decisions were
273 more frequently (28.6%) based on ambiguous statistical evidence compared to non-accelerated
274 approval (12.0%).

275 **[FIGURE 4 HERE]**

276 **Line of treatment**

277 Overall, there were 31 indications for 1st line treatment, 28 for 2nd line treatment, and nine
278 for 3rd or later line of treatment. For OS and PFS, strength of statistical evidence did not differ
279 qualitatively between lines of treatment (see Table 1). For TR, strength of statistical evidence was
280 lowest for trials supporting drugs approved for third or later line of treatment.

281 **Cancer type**

282 Overall, there were 49 indications approved for solid cancer types and 19 indications
283 approved for haematological cancer types. There was no difference in strength of statistical
284 evidence between solid and haematological cancer types (see Table 1).

285 **Approvals based on two RCTs**

286 Seven drugs were approved based on two RCTs (Table 2).

287 **[TABLE 2 HERE]**

288 Two drugs (28.6%) were approved despite both trials providing pro-null or ambiguous
289 statistical evidence. One drug (i.e., fulvestrant) was approved despite the *BFs* for both supporting
290 RCTs indicating that the treatment did not perform better than the active control on both OS (meta
291 *BF* = 0.07, meaning that the study results are about 14 times more likely to have occurred if the
292 treatment does not work than if the treatment worked) and TR (meta *BF* = 0.12, meaning that the
293 study results are about 8 times more likely to have occurred if the novel treatment does not work
294 than if it worked). Another drug (i.e., ruxolitinib phosphate) was approved based on two trials
295 providing ambiguous statistical evidence for OS improvements (meta *BF* = 0.89) and ambiguous
296 statistical evidence for PFS improvements from one trial (*BF* = 0.53)[1]. For the remaining five drugs,

[1] Note that for ruxolitinib phosphate evidence was pooled across trials with active and inactive comparators

297 one RCT with ambiguous statistical evidence was supplemented by another RCT with moderate to
298 very strong statistical evidence in favour of superiority of the treatment.

299 **Exploratory analysis**

300 **Relationship between strength of statistical evidence, effect size, and sample size**

301 For 40 effect sizes (35 for OS and five for PFS), the corresponding confidence intervals
302 indicated uncertainty regarding the effect as it included one. For these 40 ambiguous effects
303 according to the confidence intervals, 28 also had ambiguous statistical evidence according to the
304 *BF*. However, for the remaining 12, the *BF* was more informative as it indicated pro-null statistical
305 evidence (lower left and right quadrant in Figure 5). For 98 effect sizes (37 for OS and 61 for PFS), the
306 corresponding confidence intervals excluded one. For the majority of effects (*n*=89 corresponding to
307 62 indications), HR and *BF* both indicated efficacy (upper left quadrant in Figure 5). However, for 9
308 effects, the HR and corresponding confidence intervals indicated efficacy, whereas the *BF* indicated
309 ambiguous statistical evidence. There was no relationship between strength of statistical evidence
310 and sample size (r = -0.03).

311 **[FIGURE 5]**

312 Results from the exploratory qualitative analysis are presented in Table 3 and detailed in the
313 supplement (section 3).

314 **[TABLE 3]**

315 **DISCUSSION**

316 The strength of statistical evidence for pivotal RCTs supporting novel cancer drugs approved
317 by the FDA in the last two decades was substantially lower for OS, arguably the most important
318 outcome for cancer patients, compared to surrogate outcomes. Most indications (58.7%, 44/75)
319 were approved without clear statistical evidence for OS improvements. While most of these
320 indications (33/44) were supplemented by strong statistical evidence for improvements on at least
321 one surrogate endpoint (i.e., PFS/TR), uncertainties regarding OS improvements remain.

322 Strength of statistical evidence differed between approval pathways but not lines of
323 treatment or cancer types. Although few indications were approved through accelerated approval,
324 our results suggest that weaker statistical evidence is accepted for accelerated approval decisions,
325 highlighting the importance of timely post-approval studies to confirm efficacy. This need is also
326 recognized in the Consolidated Appropriations Act, 2023 (H.R. 2617), enabling the FDA to require
327 post-approval studies to be "*underway prior to granting accelerated approval*" [23]. Still, it remains
328 unclear how consistently the FDA will react to post-approval trials failing to confirm clinical benefits.
329 For example, one third of the novel cancer drugs receiving accelerated approval until 2020 but
330 subsequently failing to improve primary endpoints in their post- approval studies, remained on the

331 drug's labelling (i.e., approved under the accelerated pathway) or were converted to regular
332 approval as of July 2021 [24]. While attention to post- approval trials for indications approved via
333 accelerated pathways has increased, clear regulations and consistent action in response to them is
334 still lacking [24].

335 We also observed indications receiving non-accelerated approval with absence of statistical
336 evidence or even statistical evidence for the absence of efficacy. In these cases, approval was issued
337 based on other considerations. For example, favourable benefit-harm assessments were justified by
338 expected benefits such as improved quality of life and safety profiles. However, reporting of quality-
339 of-life-outcomes is incomplete [6] and not systematic, and a good safety profile is irrelevant in the
340 absence of efficacy. In some cases, efficacy was determined in a manner different to the pre-
341 registered protocol, for instance through eyeballing Kaplan-Meier survival curves or comparing
342 percentages. We observed one instance in which a trial that failed to demonstrate superiority was
343 re-interpreted as a non-inferiority trial. Switching between superiority and non-inferiority
344 interpretations after results are known is problematic [25,26]. It would be preferable to base drug
345 approval decisions on additional trials, which would lead to stronger statistical evidence.

346 **Strength and Limitations**

347 The current study used a comprehensive database of pre-approval trials of cancer drugs,
348 examined multiple endpoints, and used a Bayesian approach to gain novel insights into the strength
349 of evidence that allows for interpretations that are in line with clinical decision making and may
350 provide a more intuitive perspective on evidence. Nonetheless, the study also has several
351 limitations. First, we focused on the statistical evidence of RCTs supporting approval decisions. This
352 does not reflect the full complexity of the approval process and the variety of sources of evidence
353 that might be considered. Nevertheless, RCTs generally provide the strongest available evidence, as
354 other sources of evidence (such as single-trial arms) are much more difficult to interpret. Second,
355 strength of statistical evidence is only one component that is important in assessing the benefits of a
356 drug, and other factors also play a role, including the size of an effect and risk of bias. Third, we only
357 included trials conducted to support approval decisions and did not include, for instance, post-
358 approval studies. As such, our results indicate the strength of statistical evidence available to
359 support initial approval decisions and do not necessarily reflect the current strength of statistical
360 evidence.

361 **Conclusion**

362 Bayes factors offer a novel perspective and further insights into the strength of statistical
363 evidence underlying approval decisions for novel cancer drugs. Regulatory decision making could be
364 improved by using Bayes factors to distinguish between drugs with good statistical evidence, those

that lack statistical evidence (i.e., ambiguous statistical evidence), and those with statistical evidence against efficacy. We found that across the board the level of evidence for beneficial effects on overall survival is low. Some drugs were even approved without supporting statistical evidence on overall survival, progression-free survival or tumor response. These cases require a transparent and clear explanation. In many cases the statistical evidence is ambiguous calling for additional trials, before or after approval, to reduce this considerable uncertainty.

Contributions: ***M.-M.P***.: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Resources, Visualization, and Writing - original draft. ***M.L***.: Formal analysis, Software, and Writing - review & editing. ***Y.A.d.V.***: Conceptualization, Supervision, and Writing - review & editing. ***L.G.H***.: Conceptualization, Investigation, and Writing - review & editing. ***A.M.S***.: Investigation and Writing - review & editing. ***R.R.M.***: Conceptualization, Supervision, and Writing - review & editing. ***D.v.R***.: Conceptualization, Funding acquisition, Supervision, and Writing - review & editing.

Data sharing: The data that support the findings of this study are available from https://ceit-cancer.org/ and the OSF framework (https://osf.io/4uhz7).

Transparency Statement: MMP affirms that this manuscript is an honest, accurate, and transparent account of the study being reported and no important aspects of the study have been omitted. Any discrepancies from the study as originally planned and registered have been explained.

Funding: This project is funded by an NWO Vidi grant to D. van Ravenzwaaij (016.Vidi.188.001).

Patient and Public Involvement: Patients or the public were not involved in the design, or conduct, or reporting, or dissemination plans of our research.

Tables and Figures

Table 1. Available data per drug approval supported by one RCT.

	Overall		OS		PFS		TR
	N (%)	*Median BF*	*N (%)*	*Median BF*	*N (%)*	*Median BF*	*N (%)*
Overall	68 (100)		60 (100)	1.89	57 (100)	24767.8	57 (83.8)
Control group							
active	22 (32.4)	17.87	19 (31.7)	1.09	19 (33.3)	245.2	21 (36.8)
inactive	46 (67.6)	58.18	41 (68.3)	3.16	38 (66.7)	84726.9	36 (63.2)
Accelerated approval							
yes	10 (14.7)	11.58	7 (11.7)	0.45	7 (12.3)	426.45	7 (12.3)
no	58 (85.3)	41.09	53 (88.3)	2.02	50 (87.7)	26629.25	50 (87.7)
Line of treatment							
first	31 (45.6)	17.87	30 (50.0)	2.00	23 (40.4)	28490.71	24 (42.1)
second	28 (41.2)	83.57	22 (36.7)	1.73	26 (45.6)	2631.38	26 (45.6)
≥ third	9 (13.2)	5.00	8 (13.3)	1.70	8 (14.0)	17761.5	7 (12.3)
Cancer type							
solid	49 (72.1)	32.03	46 (76.7)	2.12	46 (80.7)	26629.25	41 (71.9)
haematological	19 (27.9)	16.28	14 (23.3)	1.04	11 (19.3)	20494.97	16 (28.1)

Table 2. *BFs corresponding to the individual trials and the meta-analytic BF for all drugs with more than one RCT. BFs are presented per outcome.* Yellow: *BF*<⅓; Light green: ⅓ <*BF*<3, Dark green: 3<*BF*<10, Turquoise: 10<*BF*<30; Dark blue: *BF*>30

Drug	**Outcome**	**Strength of statistical evidence (*BF*)**		
		RCT1	**RCT2**	**Combined**
fulvestrant	OS	0.11	0.08	0.07
	PFS	-	-	-
	TR	0.28	0.09	0.33
ruxolitinib phosphate	OS	0.95	0.81	0.89
	PFS	0.53	-	-
	TR	-	-	-
trastuzumab emtansine	OS	0.34	77.48	40.26
	PFS	52279.93	-	-
	TR	0.26	78.61	76.30
bevacizumab	OS	2.74	768.15	2060.87
	PFS	20.34	$7.81*10^{8}$	$1.38*10^{9}$
	TR	2.70	5.89	25.6
trifluridine tipiracil	OS	30.20	1865.24	29179.5
	PFS	$1.79*10^{15}$	-	-
	TR	0.04	0.05	0.71
binimetinib	OS	-	79.94	
	PFS	2.62	1677.46	4286.21
	TR	4.49	2.11	45.4
sorafenib tosylate	OS		4.48	
	PFS	697.30	$5.96*10^{10}$	$1.20**10^{10}$
	TR		0.90	

Table 3. Reasons for drug approval mentioned for the seven drugs without statistical evidence for OS, PFS, or TR improvements. These themes were extracted via an exploratory qualitative thematic analysis.

Theme	Subthemes
Expected benefits over (non)-existing treatments due to…	poor prognosis (despite available treatments)
	few or highly toxic available treatments
	prior failed treatments
	different administration (e.g., oral medication)
Tolerability	limited crossover between treatment arms
	low discontinuation rates
Positive interpretation of limited statistical evidence	meaningful clinical difference (e.g., difference being statistically not significant but effect size deemed clinically meaningful)
	convincing interim analyses (e.g., very small *p*-value interpreted as robust statistical evidence)
	descriptive interpretation of premature survival analysis (e.g., eyeballing Kaplan-Meier survival curves or comparing percentages)
	non-inferiority to control group (e.g., failed superiority trial interpreted as successful non-inferiority trial)
Primary endpoints different from OS, PFS, or TR	regulatory precedents (e.g., other drugs in the same group were approved based on this endpoint)
	experience (e.g., the alternative endpoint was known to relate to OS or PFS)
Use of non-randomized evidence	control group deemed unethical

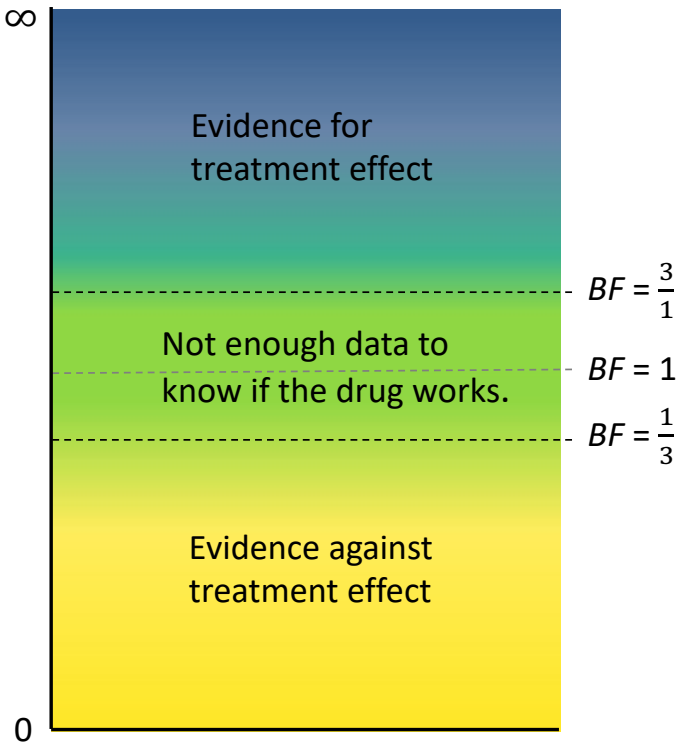

Figure 1. *Visual representation of the continuous properties of the Bayes Factor in the context of treatment evaluation. Bayes Factors above 3 indicate evidence in favour of the treatment effect. Bayes Factors below 1/3 indicate evidence against the treatment effect.*

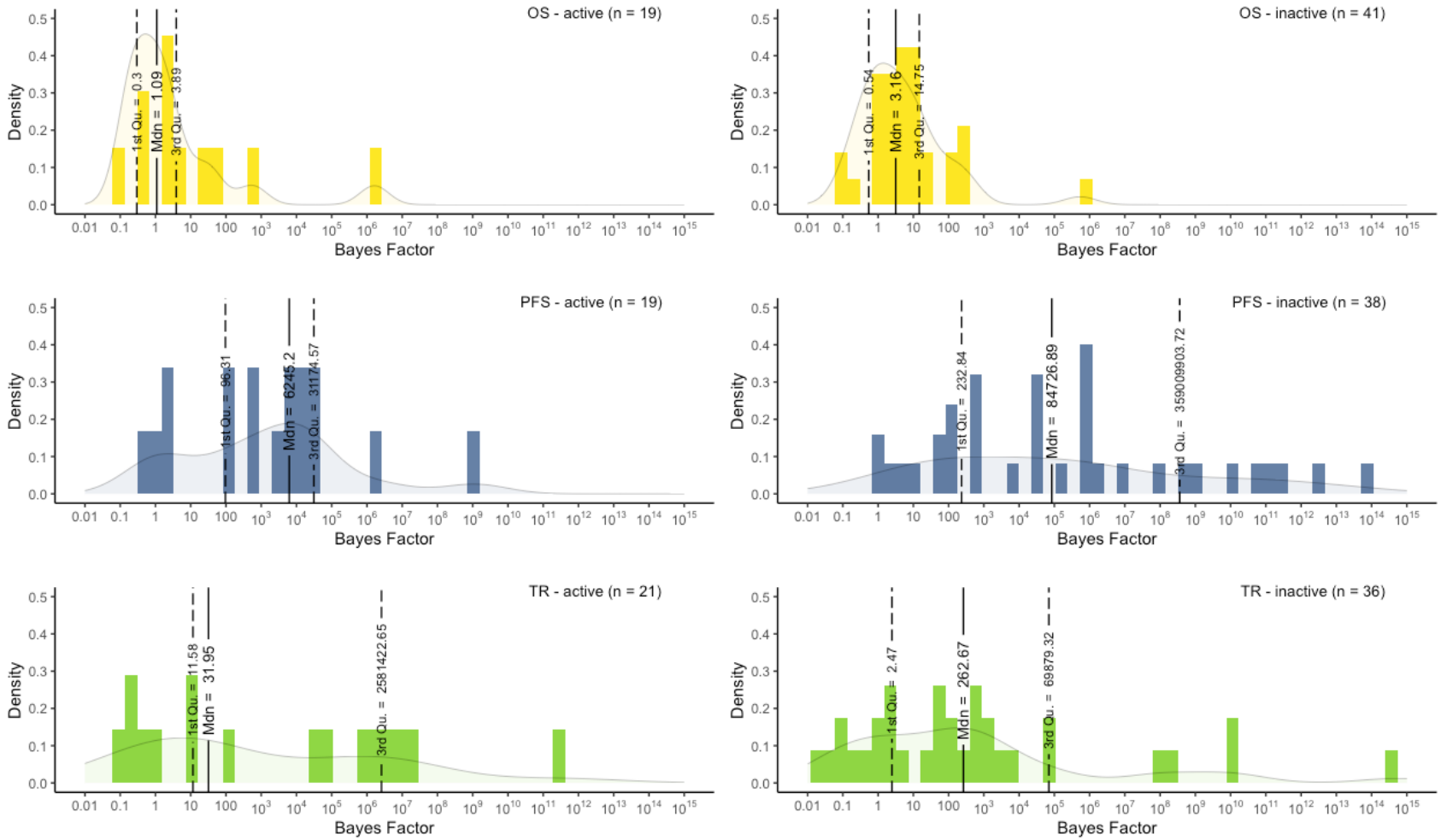

Figure 2. *Histograms and density plots illustrating the distribution of BFs for the three possible endpoints and two types of comparators. Medians and first and third quartiles are presented in vertical lines. Note that for PFS-active 2 effects, for PFS-inactive 2 effects, TR-active 1 effect, and TR-inactice 3 effects were cut off.*

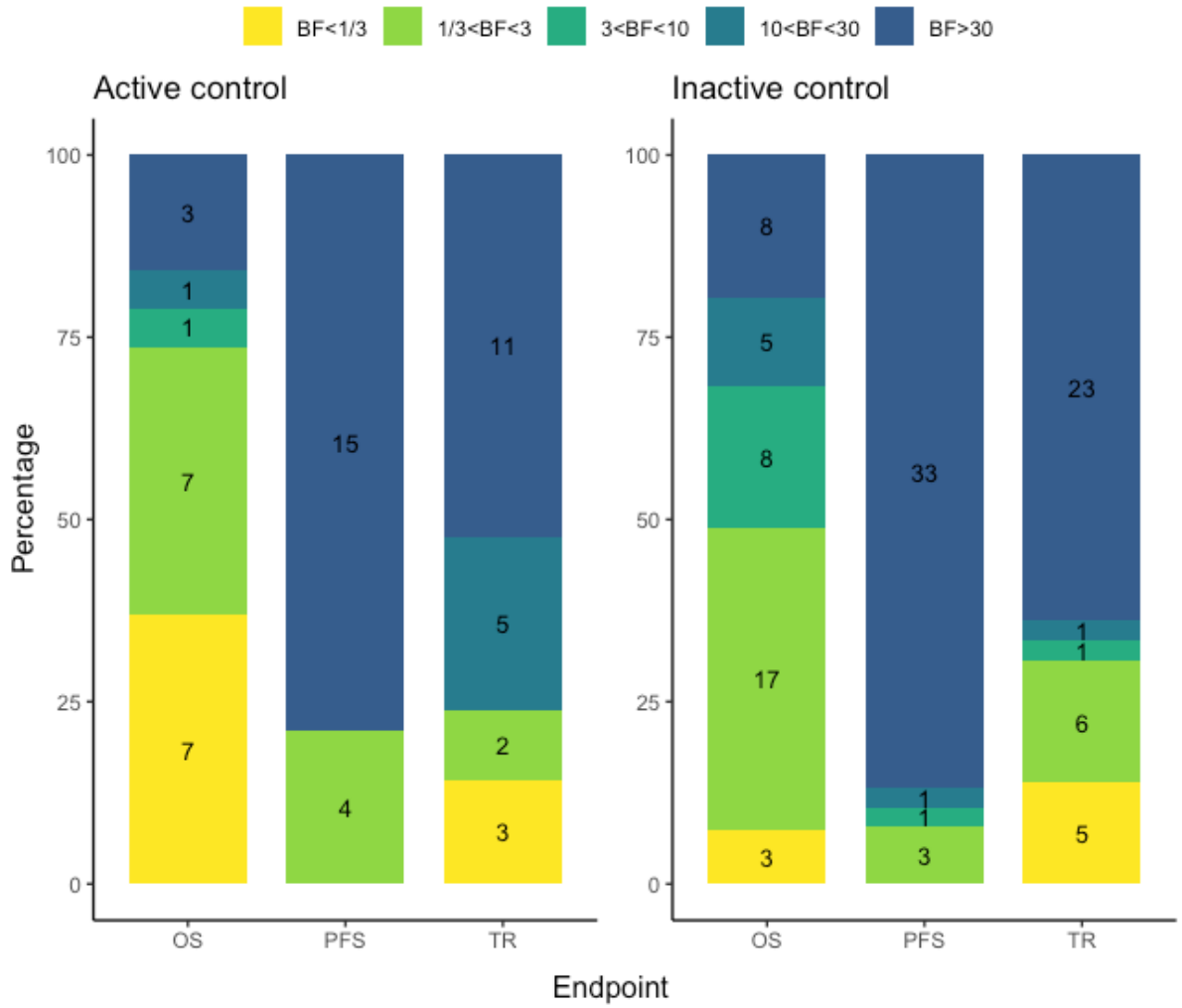

Figure 2. *Bar plot illustrating the proportion of trials supported by pro-null evidence (BF<⅓), ambiguous evidence (⅓<BF<3), moderate pro-alternative evidence (3<BF<10), strong pro-alternative evidence (10<BF<30) and very strong pro-alternative evidence per endpoint and control condition. Counts are presented per category. Note that indications based on multiple trials are not included in this figure.*

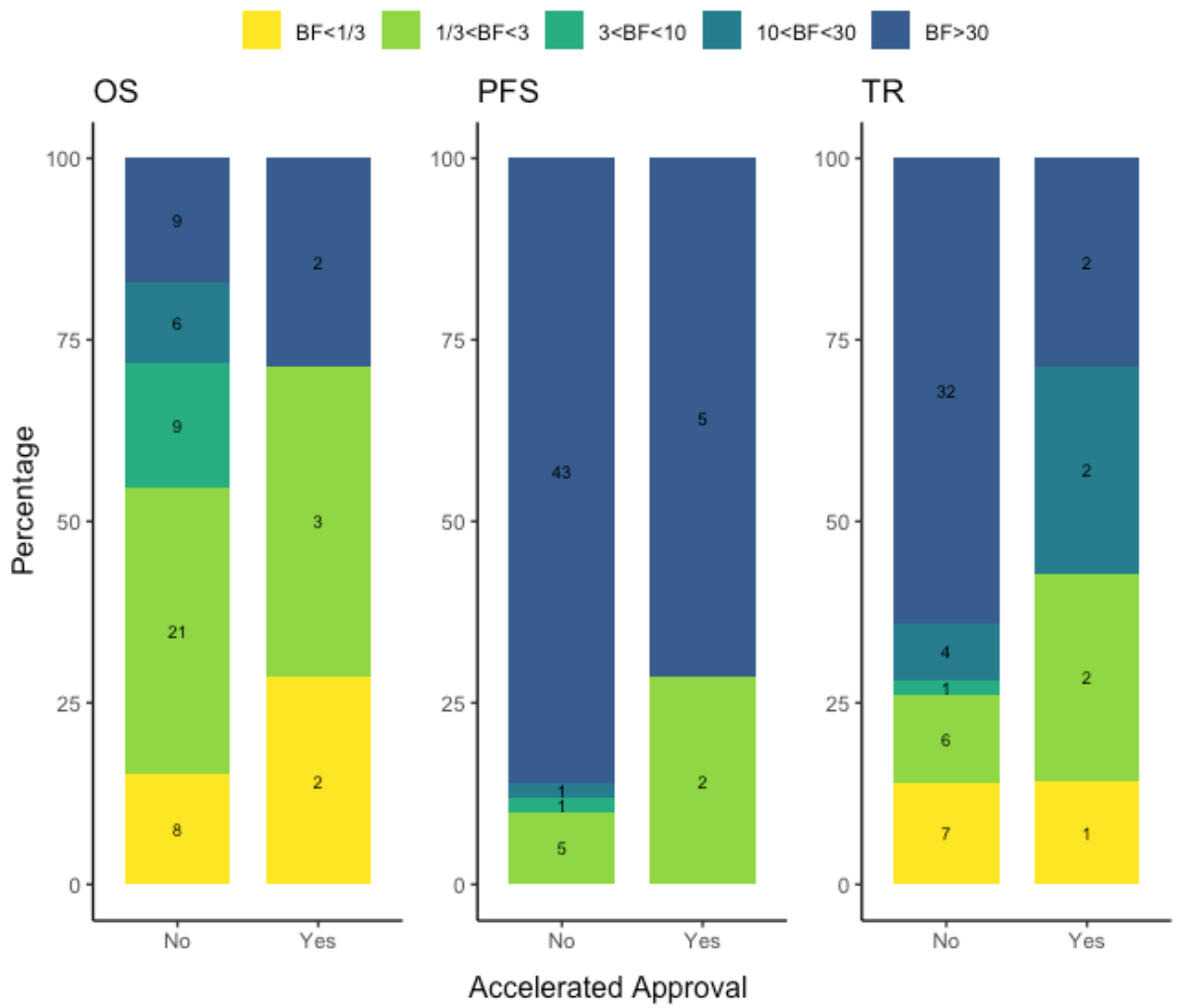

Figure 3. *Bar plot illustrating the proportion of trials supported by pro-null evidence (BF<⅓), ambiguous evidence (⅓<BF<3), moderate pro-alternative evidence (3<BF<10), strong pro-alternative evidence (10<BF<30) and very strong pro-alternative evidence per endpoint and approval pathway. Note that indications based on multiple trials are not included in this figure*

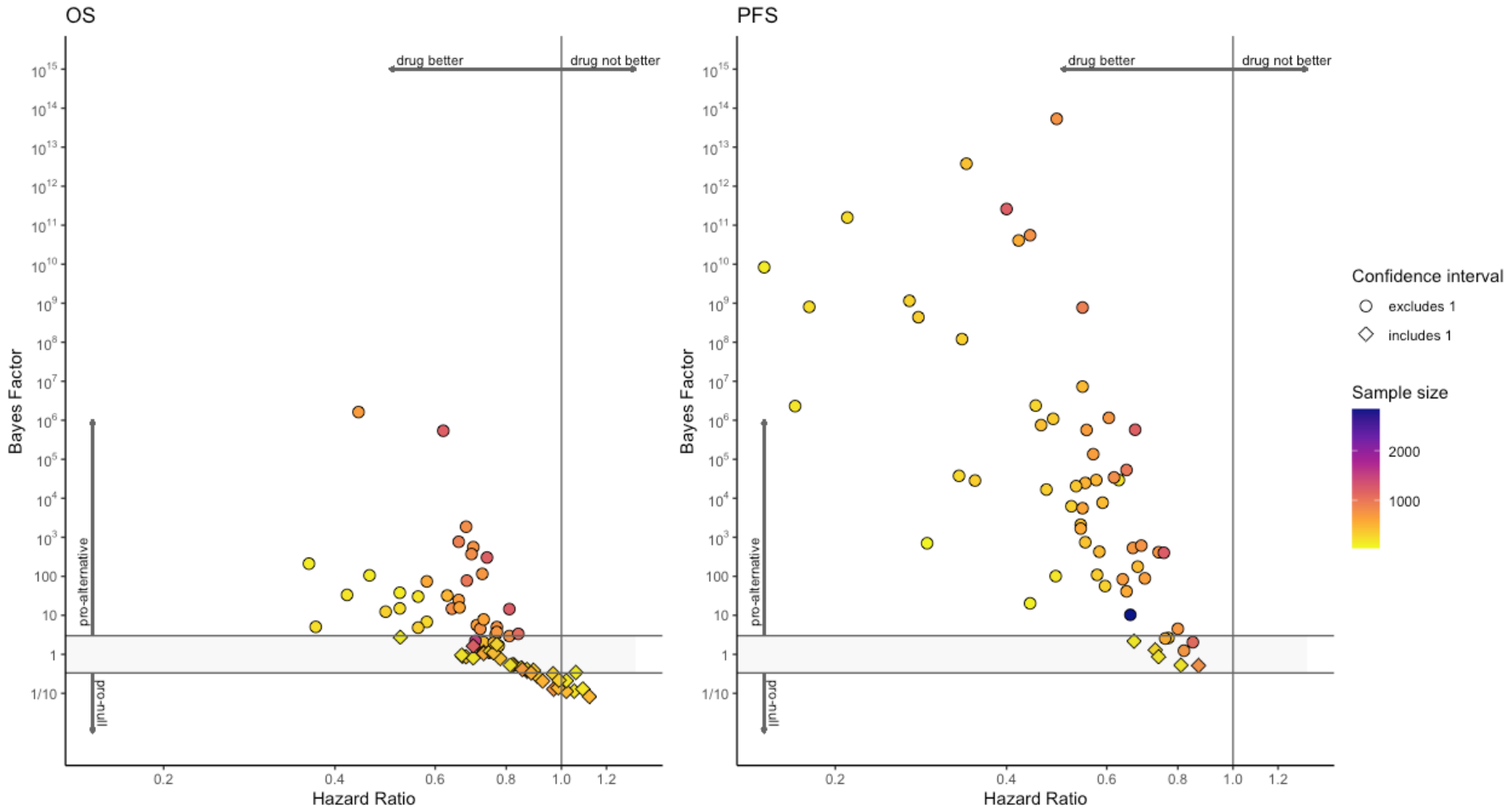

Figure 4. *BFs plotted against HR for the 68 indications with one RCT. Shapes indicate whether or not the confidence interval of the HR included 1.*

References

1 Ladanie A, Schmitt AM, Speich B, *et al.* Clinical Trial Evidence Supporting US Food and Drug Administration Approval of Novel Cancer Therapies Between 2000 and 2016. *JAMA Netw Open* 2020;**3**:e2024406–e2024406. doi:10.1001/jamanetworkopen.2020.24406

2 Gyawali B, Hey SP, Kesselheim AS. Assessment of the Clinical Benefit of Cancer Drugs Receiving Accelerated Approval. *JAMA Intern Med* 2019;**179**:906. doi:10.1001/jamainternmed.2019.0462

3 Salas-Vega S, Iliopoulos O, Mossialos E. Assessment of Overall Survival, Quality of Life, and Safety Benefits Associated With New Cancer Medicines. *JAMA Oncol* 2017;**3**:382. doi:10.1001/jamaoncol.2016.4166

4 Vokinger KN, Kesselheim AS. Characteristics of trials and regulatory pathways leading to US approval of innovative vs. non-innovative oncology drugs. *Health Policy* 2019;**123**:721–7. doi:10.1016/j.healthpol.2019.06.002

5 Michaeli DT, Michaeli T. Overall Survival, Progression-Free Survival, and Tumor Response Benefit Supporting Initial US Food and Drug Administration Approval and Indication Extension of New Cancer Drugs, 2003-2021. *J Clin Oncol Off J Am Soc Clin Oncol* 2022;:JCO2200535. doi:10.1200/JCO.22.00535

6 Gloy V, Schmitt AM, Düblin P, *et al.* The Evidence Base of US Food and Drug Administration Approvals of Novel Cancer Therapies from 2000 to 2020. *Int J Cancer* Published Online First: 2023. doi:10.1002/ijc.34473

7 U.S. Food and Drug Administration. Guidance for Industry: E9 Statistical Principles for Clinical Trials. 1998.

8 Kemp R, Prasad V. Surrogate endpoints in oncology: when are they acceptable for regulatory and clinical decisions, and are they currently overused? *BMC Med* 2017;**15**:134. doi:10.1186/s12916-017-0902-9

9 Pignatti F, Jonsson B, Blumenthal G, *et al.* Assessment of benefits and risks in development of targeted therapies for cancer — The view of regulatory authorities. *Mol Oncol* 2015;**9**:1034–41. doi:https://doi.org/10.1016/j.molonc.2014.10.003

10 Prasad V, Kim C, Burotto M, *et al.* The Strength of Association Between Surrogate End Points and Survival in Oncology: A Systematic Review of Trial-Level Meta-analyses. *JAMA Intern Med* 2015;**175**:1389–98. doi:10.1001/jamainternmed.2015.2829

11 Gronau QF, Ly A, Wagenmakers E-J. Informed Bayesian T-Tests. *Am Stat* 2019;:1–14. doi:10.1080/00031305.2018.1562983

12 Rouder JN, Speckman PL, Sun D, *et al.* Bayesian t tests for accepting and rejecting the null hypothesis. *Psychon Bull Rev* 2009;**16**:225–37. doi:10.3758/PBR.16.2.225

13 Jeffreys H. *Theory of probability*. Oxford: : Oxford University Press 1961.

14 Ladanie A, Speich B, Naudet F, *et al.* The Comparative Effectiveness of Innovative Treatments for Cancer (CEIT-Cancer) project: Rationale and design of the database and the collection of evidence available at approval of novel drugs. *Trials* 2018;**19**:505. doi:10.1186/s13063-018-2877-z

15 R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: : R Foundation for Statistical Computing 2021. https://www.R-project.org/

16 Linde M, van Ravenzwaaij D, Tendeiro JN. *baymedr: Computation of Bayes Factors for Common Biomedical Designs*. 2022. https://github.com/maxlinde/baymedr

17 Linde M, Tendeiro JN, Ravenzwaaij D van. Bayes Factors for Two-group Comparisons in Cox Regression. 2022;:2022.11.02.22281762. doi:10.1101/2022.11.02.22281762

18 Morey RD, Rouder JN, Jamil T, *et al.* BayesFactor: Computation of Bayes Factors for Common Designs. 2022.https://CRAN.R-project.org/package=BayesFactor (accessed 19 Jul 2022).

19 Jamil T, Ly A, Morey RD, *et al.* Default “Gunel and Dickey” Bayes factors for contingency tables. *Behav Res Methods* 2017;**49**:638–52. doi:10.3758/s13428-016-0739-8

20 Heck DW, Gronau QF, Wagenmakers E-J. metaBMA: Bayesian model averaging for random and

fixed effects meta-analysis. Retrieved Doi. 2017.https://cran.r-project.org/web/packages/metaBMA/metaBMA.pdf (accessed 20 Mar 2021).

21 Gronau QF, Heck DW, Berkhout SW, *et al.* A Primer on Bayesian Model-Averaged Meta-Analysis. *Adv Methods Pract Psychol Sci* 2021;**4**:25152459211031256. doi:10.1177/25152459211031256

22 Lee MD, Wagenmakers E-J. *Bayesian cognitive modeling: A practical course*. Cambridge university press 2014.

23 New FDA Reform Legislation: Congress Gifts a "FDORA" for the Holidays. 2023.https://www.ropesgray.com/en/newsroom/alerts/2023/01/new-fda-reform-legislation-congress-gifts-a-fdora-for-the-holidays (accessed 16 Feb 2023).

24 Gyawali B, Rome BN, Kesselheim AS. Regulatory and clinical consequences of negative confirmatory trials of accelerated approval cancer drugs: retrospective observational study. *BMJ* 2021;**374**:n1959. doi:10.1136/bmj.n1959

25 U.S. Food and Drug Administration, Center for Drug Evaluation and Research. Non-Inferiority Clinical Trials to Establish Effectiveness Guidance for Industry. 2016.

26 Committee for Proprietary Medicinal Products. Points to consider on switching between superiority and non-inferiority. *Br J Clin Pharmacol* 2001;**52**:223–8. doi:10.1046/j.0306-5251.2001.01397-3.x