

26 **ABSTRACT**

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

29 **Study Design and Setting:** We used data on overall survival (OS), progression-free survival (PFS), and
30 tumour response (TR) for novel cancer drugs approved for the first time by the FDA between January
31 2000 and December 2020. We assessed strength of statistical evidence by calculating Bayes Factors
32 (*BFs*) for all available endpoints, and we pooled evidence using Bayesian fixed-effect meta-analysis
33 for indications approved based on two RCTs. Strength of statistical evidence was compared between
34 endpoints, approval pathways, lines of treatment, and types of cancer.

35 **Results:** We analysed the available data from 82 RCTs corresponding to 68 indications supported by
36 a single RCT and seven indications supported by two RCTs. Median strength of statistical evidence
37 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*10⁶)
38 and TR (*BF* = 113.9; IQR 3.0-547,100). Overall, 44 indications (58.7%) were approved without clear
39 statistical evidence for OS improvements and seven indications (9.3%) were approved without
40 statistical evidence for improvements on any endpoint. Strength of statistical evidence was lower for
41 accelerated approval compared to non-accelerated approval across all three endpoints. No
42 meaningful differences were observed for line of treatment and cancer type.

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

45 **Conclusion:** *BFs* offer novel insights into the strength of statistical evidence underlying cancer drug
46 approvals. Most novel cancer drugs lack strong statistical evidence that they improve OS, and a few
47 lack statistical evidence for efficacy altogether. These cases require a transparent and clear
48 explanation. When evidence is ambiguous, additional post-marketing trials could reduce
49 uncertainty.

50 **Keywords:** cancer drugs, evidence, regulatory approval, Bayes Factor, overall survival

51

BACKGROUND

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67

68

69

70

71

72

73

74

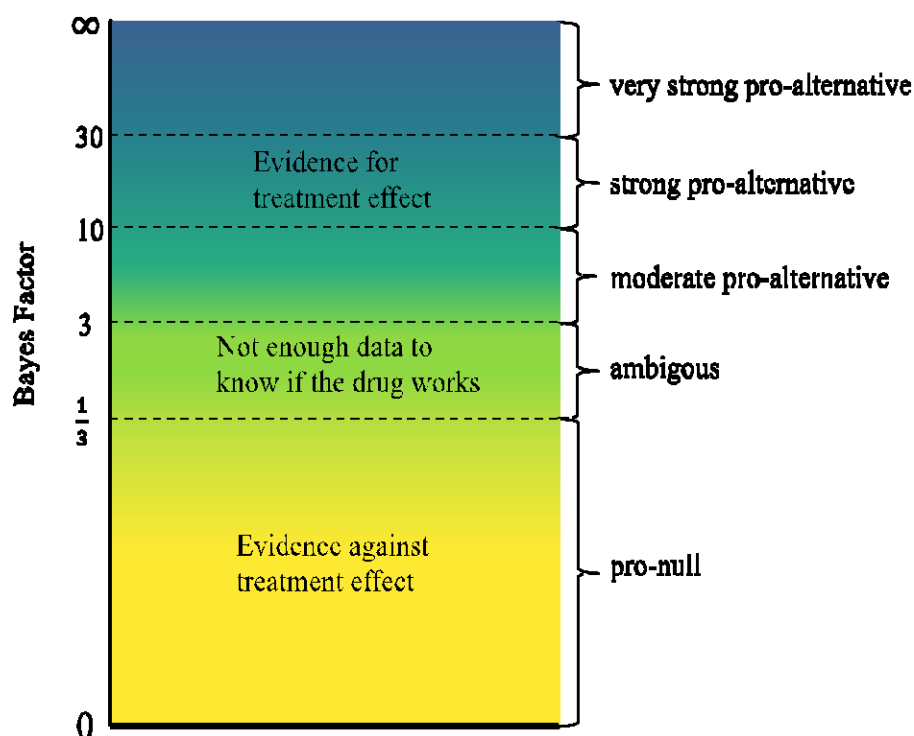
75

76

For a new cancer drug to be marketed in the U.S., it needs to be endorsed by the Food and Drug Administration (FDA). While many aspects of a drug's profile are considered in the approval process, the statistical evaluation of efficacy plays a central role [1]. Cancer drugs are frequently approved based on limited evidence, which increases uncertainty in clinical decision making [2–7]. Despite FDA guidelines suggesting that substantial evidence for efficacy based on two convincing trials should be provided [8], approval for novel cancer drugs between 2000 and 2020 was typically based on a single pivotal (i.e., efficacy determining) trial [2,7]. The understanding that this leads to a reduced level of statistical evidence for efficacy is implicit. Previous work has considered the strength of evidence indirectly considering effect sizes, number and kinds of trials, or qualitative evidence [2,7]. Explicit quantification of the statistical strength of evidence is missing. Another complication arises from the FDA permitting surrogate endpoints, that is, outcomes that are no “direct measurement[s] of clinical benefit but [] known to predict clinical benefit” (p. 16 [9]) such as overall survival (OS) but are easier or faster to measure [10]. However, evidence that surrogate endpoints, like progression-free survival (PFS) and tumour response (TR), predict overall survival in oncology is limited [10–12], limiting the quality of the evidence available at the time of approval. Questions have been raised whether the current statistical evidence for cancer drugs meets the FDA's requirement for showing “meaningful therapeutic benefits” [3].

Explicit quantification of statistical strength of evidence for efficacy at the time of approval can be achieved with Bayes Factors (*BFs*). Technically, *BFs* compare the likelihood of the observed data under the null hypothesis (i.e., novel treatment does not improve outcomes) to the likelihood of the observed data under an alternative hypothesis (e.g., novel treatment improves outcomes; one-sided alternative hypothesis) [13–15]. The resulting ratio provides a relative measure of statistical evidence for or against competing hypotheses (see Figure 1). For example, a *BF* of 1 indicates 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 *BF* of 10 indicates that the

77 observed study results are ten times more likely to have occurred given that the novel treatment
78 improves OS than if it does not improve OS, whereas a *BF* of 0.1 indicates that the observed study
79 results are ten times more likely to have occurred given that the novel treatment does not work
80 than if it does work.



81
82 Figure 1. Visual representation of the continuous properties and thresholds of the Bayes Factor in the
83 context of treatment evaluation. Bayes Factors above 3 indicate evidence in favour of the treatment
84 effect. Bayes Factors below $1/3$ indicate evidence against the treatment effect. Note: These
85 thresholds have been proposed by methodological researchers and are not clinically informed
86 [15,24].

87 Using the *BF*, we offer a different perspective on statistical evidence that may be more
88 intuitive than the traditional focus on statistical significance in frequentist frameworks. We take a
89 perspective like that of a clinician who wants to know if the results of a diagnostic test (here, the
90 approval trials) are more likely under a working diagnosis than under a differential diagnosis, which
91 is reflected by the *BF*. Moreover, using *BFs*, unlike traditional Frequentist testing, enables one to

92 differentiate between absence of evidence (i.e., ambiguous evidence) and evidence of absence (i.e.,
93 pro-null evidence). This project has two aims. First, we aim to quantify and describe the strength of
94 statistical evidence for efficacy associated with novel¹ cancer drugs approved between 2000 and
95 2020 using *BFs*. Our second aim is to contrast strength of statistical evidence for efficacy between
96 endpoints, approval types, lines of treatment, and type of cancer, as standards for what constitutes
97 “sufficient” strength of statistical evidence for beneficial effects (and against harmful effects) at the
98 time of approval might differ. For example, one might accept more uncertainty and lower strength
99 of statistical evidence for 3rd or 4th line treatments than for 1st or 2nd line treatments. Previous
100 reports also suggest that novel drugs for haematological cancers are more likely to be approved
101 based on single-arm trials and surrogate endpoints compared to solid cancers [2].

102 METHOD

103 Data and Registration

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

107 FDA Data

108 We used data from the CEIT-Cancer project (details provided elsewhere [16,17]). In short,
109 novel drugs and biological therapies receiving first approval for the treatment of any malignant
110 diseases between January 2000 and December 2020 and the corresponding FDA reviews (available
111 at drugs@FDA) were identified. For all RCTs evaluating the drug in the approved indication
112 (regardless of whether the trial was described as pivotal or not), the following data were extracted:
113 hazard ratios (HR) and associated 95% confidence intervals for OS and/or PFS; median OS and/or
114 PFS; number of events for OS, PFS and/or TR; sample size; line of treatment; approval pathway (i.e.,
115 priority review, orphan design, accelerated approval, and breakthrough therapy designation); type

¹ Novel meaning first FDA approvals as opposed to label extensions or new indications of previously approved drugs.

116 of cancer; type of control; and type of blinding. The original data set also included single-arm
117 clinical trials that were explicitly described as pivotal. However, here we only considered data from
118 RCTs as single-arm trials do not provide comparative treatment outcomes.

119 **Data Analysis**

120 We used R version 4.1.2 [18]. We calculated *BFs* for all available endpoints per RCT. For OS
121 and PFS, we used the available summary statistics (i.e, HR, confidence interval of HR, sample size,
122 and number of events per group) to conduct a Bayesian Cox regression using the “baymedr” R
123 package [19]. We used a standard normal distribution that was truncated at 0 as the prior for beta
124 under the alternative hypothesis. For TR outcomes, *BFs* for chi-square tests were calculated using
125 the “BayesFactor” R package [20]. Under the null hypothesis, the prior for the proportion of TR was
126 a joint uniform distribution ranging from 0 to 1. Under the alternative hypothesis, the prior for the
127 proportion of TR was a uniform distribution ranging from 0 to 1 for each group independently [21].
128 See supplement (section 1) for further details. For approvals based on two RCTs, we pooled available
129 outcomes for OS and PFS via fixed-effect meta-analysis using the “metaBMA” R package [22]. Under
130 the alternative hypothesis, the prior for the treatment effect was a default Cauchy distribution that
131 was truncated at, with a location parameter of 0 and a scale parameter of $\frac{1}{\sqrt{2}}$ [23]. For TR, we pooled
132 the number of events and number of participants and calculated corresponding chi-square tests
133 using the “BayesFactor” R package with the same specifications as for single trial results. To describe
134 the strength of statistical evidence we adopted standard thresholds (see Figure 1).

135 The interpretation of pro-null *BFs* depends on the control group. Trials were classified as
136 inactively controlled if the control group received a placebo or no treatment even if all study
137 patients received treatment as usual (e.g., chemotherapy). Trials were classified as actively
138 controlled if the control group received another active treatment different from the experimental
139 group (e.g., an established treatment). In trials with inactive control groups, a pro-null *BF* means that
140 there is evidence that the drug performs comparable to or worse than placebo or no treatment,
141 indicating no efficacy of the novel drug. In trials with an active control group, interpreting pro-null

142 *BFs* is more difficult. Assuming that the active control group receives an effective therapy, a pro-null
143 *BF* can mean that the novel drug is ineffective, or just that it is not *more* effective than an
144 established effective treatment. Therefore, we discuss evidence per control condition whenever
145 possible. For the subgroup analysis of approval pathway, line of treatment, and type of cancer, we
146 did not separate results by control group as these subgroups were too small to be split further. We
147 describe results separately for drugs supported by a single RCT versus drugs supported by two RCTs,
148 as the two RCTs supporting a single drug may use different control groups (i.e., one using an inactive
149 control group and one using an active control group). This decision was not pre-registered.

150 **Exploratory data analysis**

151 We conducted four not pre-registered analyses, exploring: (1) the relationship between *BFs*
152 and effect sizes, (2) the relationship between *BFs* and sample size, and (3) the qualitative reasoning
153 behind endorsement decisions approved based on pro-null or ambiguous statistical evidence.
154 Rationales and details are provided in the supplement (section 3), and (4) in response to a reviewer
155 comment we examined the median strength of statistical evidence for primary endpoints and (5) the
156 relationship between evidential strength of OS and PFS, which is reported in the supplement.

157 **RESULTS**

158 The dataset contained data on 145 novel cancer drugs for 156 indications based on unique
159 186 trials. Of these, 75 indications (48.1%) received FDA approval without supporting evidence from
160 RCTs, 70 indications (44.9%) were supported by one, and 11 (7.1%) indications were supported by
161 two RCTs. Summary data were unavailable for 10 RCTs. Consequently, we analysed 82 RCTs
162 supporting approval for 75 indications of 75 novel cancer drugs, 68 (90.7%) supported by a single
163 RCT and 7 (8.5%) supported by two RCTs (see Figure S1). Of the 82 included trials, 72 trials assessed
164 OS (87.8%), 66 trials assessed PFS (80.5%), and 68 trials assessed TR (82.9%). Overall, 52 trials
165 provided data for all three endpoints (63.4%), 20 trials data for two endpoints (24.4%), and 10 trials
166 for only one endpoint (12.2%). Trial characteristics and individual *BFs* per trial and endpoint are
167 presented in the supplement (Table S1 and S2).

168 **Approval decisions based on one RCT**

169 Median evidential strength for drugs approved with one supportive RCT is provided in Table
170 1, and the distribution of *BFs* per endpoint and control group are presented in Figure 2.

171 **Endpoints**

172 Statistical evidence for a beneficial effect on at least one endpoint was moderate for five out
173 of 68 (7.4%), and (very) strong for 49 out of 68 (86.8%) indications. The median *BF* for OS was 1.9
174 (IQR 0.5-14.5), for PFS 24,767.8 (IQR 109.0-7.3*10⁶) and for TR 113.9 (IQR 3.0 - 547,100).

175 **Active control groups.**

176 Trials with active control groups ($n=22$) had a median *BF* of 17.9 (IQR 1.0 - 12,220.0) across
177 all three endpoints. Three (13.6%) indications lacked (very) strong statistical evidence for benefits on
178 any endpoint, while five indications (22.7%) were supported by (very) strong statistical evidence for
179 benefits across all three endpoints.

180 Of the 19 indications with data available for OS, two (10.5%) were approved with pro-null or
181 ambiguous statistical evidence for OS improvements without (very) strong evidence for
182 improvements on surrogate endpoints. Additionally, 12 indications (63.2%) were approved based on
183 pro-null or ambiguous statistical evidence for OS improvements but with (very) strong statistical
184 evidence for improvements on at least one surrogate outcome. One indication (5.3%) was approved
185 based on moderate statistical evidence for OS improvements and (very) strong evidence for
186 improvements on the surrogate outcomes. The remaining four (21.1%) indications were approved
187 based on (very) strong statistical evidence for OS improvements.

188 For PFS or TR outcomes, the majority of trials indicated (very) strong statistical evidence for
189 better treatment outcomes in the experimental compared to the control group ($n = 15$, 79.0% and n
190 = 16, 76.2% respectively; see also Figure 3). One drug (i.e., dasatinib) was approved based on
191 ambiguous statistical evidence for TR only.

192

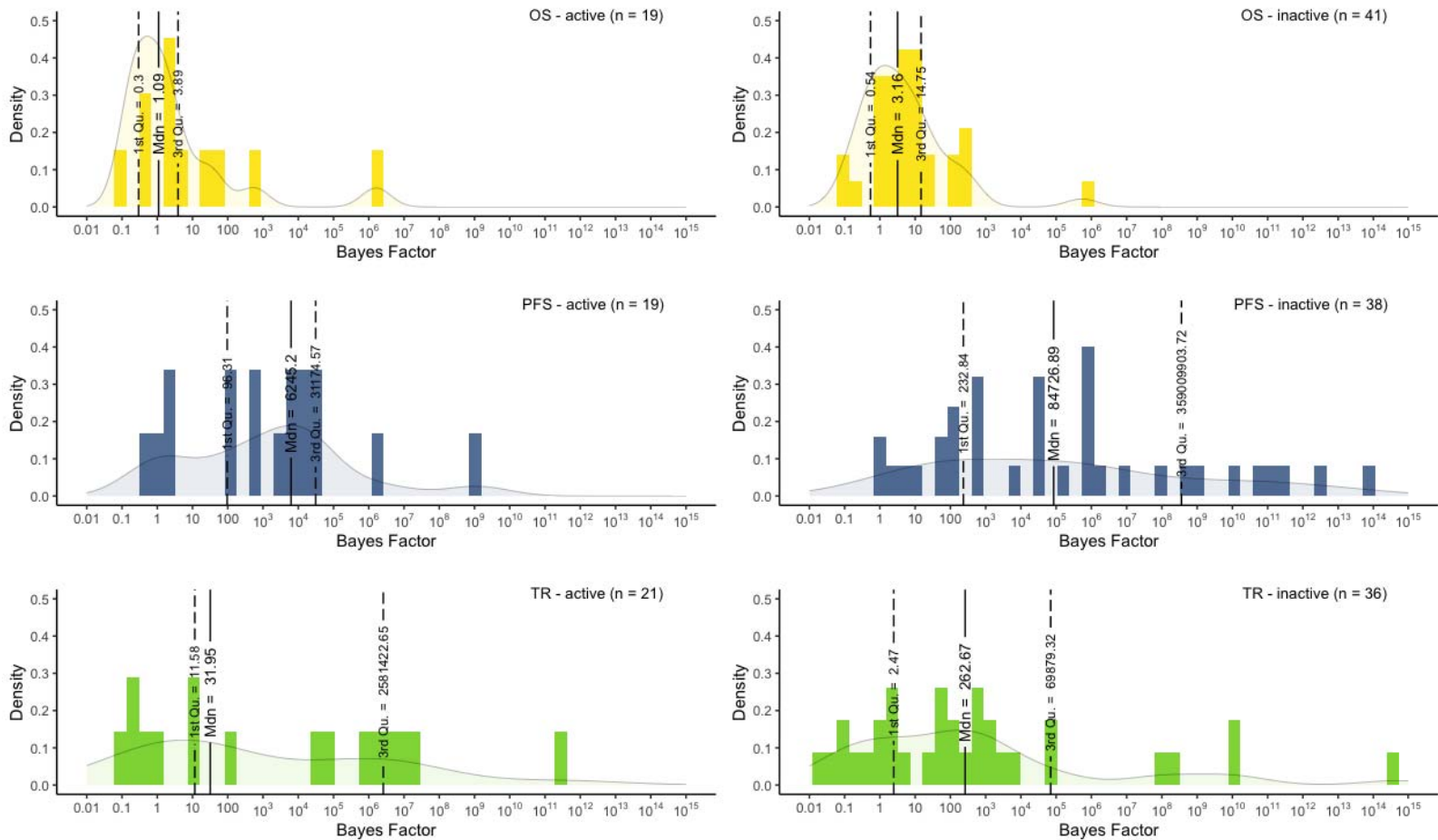
Table 1. Descriptives for drug approvals supported by one RCT.

	<i>Overall</i>		<i>OS</i>		<i>PFS</i>		<i>TR</i>	
	<i>N (%)</i>	<i>Median BF (IQR)</i>	<i>N (%)</i>	<i>Median BF [IQR]</i>	<i>N (%)</i>	<i>Median BF [IQR]</i>	<i>N (%)</i>	<i>Median BF [IQR]</i>
Overall	68 (100)		60 (100)	1.9 [0.5; 14.5]	57 (100)	24767.8 [109; 7287000]	57 (83.8)	114 [3; 547100]
Control group								
active	22 (32.4)	17.87 [1.0; 12220]	19 (31.7)	1.1 [0.3; 3.9]	19 (33.3)	245.2 [96.0; 31180]	21 (36.8)	32.0 [12.0; 2581000]
inactive	46 (67.6)	58.18 [2.0; 47080]	41 (68.3)	3.2 [0.5; 14.7]	38 (66.7)	84726.9 [233.0; 3.5*10 ⁸]	36 (63.2)	263 [2.0; 69880]
Accelerated approval								
yes	10 (14.7)	11.58 [0; 134]	7 (11.7)	0.5 [0.3; 16.9]	7 (12.3)	426.5 [52.0; 1165904]	7 (12.3)	32.0 [12.0; 2581000]
no	58 (85.3)	41.09 [1.0; 34220]	53 (88.3)	2.0 [0.8; 14.4]	50 (87.7)	26629.3 [233.0; 3.5*10 ⁸]	50 (87.7)	262.7 [2.0; 69880]
Line of treatment								
first	31 (45.6)	17.9 [2.0; 28900]	30 (50.0)	2.0 [0.9; 13.0]	23 (40.4)	28490.7 [105.0; 5.8 *10 ⁸]	24 (42.1)	621.5 [9.0; 5366000]
second	28 (41.2)	83.6 [1.0; 64850]	22 (36.7)	1.7 [0.5; 13.9]	26 (45.6)	2631.4 [418.0; 6060000]	26 (45.6)	123.9 [9.0; 435200]
≥ third	9 (13.2)	5.00 [0.0; 4310]	8 (13.3)	1.7 [0.4; 12.1]	8 (14.0)	17761.5 [43.0; 2.1*10 ⁹]	7 (12.3)	1.1 [0.43; 1199]
Cancer type								
solid	49 (72.1)	32.03 [1.0; 29110]	46 (76.7)	2.1 [0.6; 15.0]	46 (80.7)	26629.3 [233; 6041000]	41 (71.9)	107.1 [2.0; 99710]
haematological	19 (27.9)	16.28 [1.0; 22690]	14 (23.3)	1.0 [0.3; 4.7]	11 (19.3)	20495.0 [72.0; 4.0 *10 ⁸]	16 (28.1)	1069.38 [15.0; 1426000]

Note: The Overall column describes the values across all three endpoints. This means that a single trial can be included multiple times here (once for each available outcome).

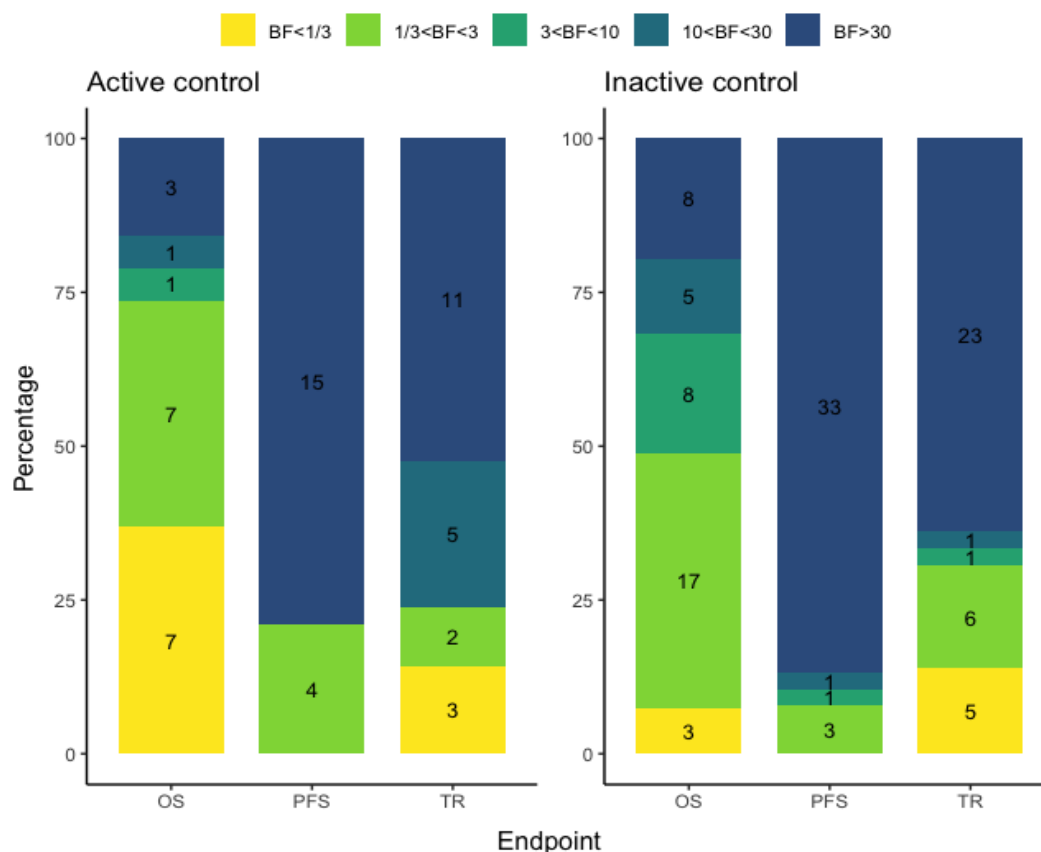
193

194



195 Figure 2. Histograms and density plots illustrating the distribution of BFs for the three possible endpoints and two types of comparators. Medians and first
 196 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-inactive 3 effects
 197 were cut off.

198



199

200 Figure 3. Bar plot illustrating the proportion of trials supported by pro-null evidence ($BF < \frac{1}{2}$),
 201 ambiguous evidence ($\frac{1}{3} < BF < 3$), moderate pro-alternative evidence ($3 < BF < 10$), strong pro-alternative
 202 evidence ($10 < BF < 30$) and very strong pro-alternative evidence per endpoint and control condition.
 203 Counts are presented per category. Note that indications based on multiple trials are not included in
 204 this figure.

205 **Inactive control groups.**

206 Trials with inactive control groups (placebo: $n = 29$; supportive care: 17) had a median BF of
 207 58.2 (IQR 2 – 47,080.0) across all three endpoints. Two (4.3%) indications lacked (very) strong
 208 evidence for benefits on any endpoint (questioning the efficacy of the drug), while seven (17.1%)
 209 were supported by (very) strong statistical evidence for benefits across all three endpoints.

210 Of 41 indications with data available for OS, two (4.9%) were approved with pro-null or
 211 ambiguous statistical evidence for improvements of OS and all other endpoints. Additionally, 18

212 drugs (43.9%) were approved based on pro-null or ambiguous statistical evidence for OS
213 improvements but with (very) strong statistical evidence for improvements on at least one surrogate
214 outcome. Eight (19.5%) indications were approved based on moderate statistical evidence for OS
215 improvements, of which five were supplemented with very strong statistical evidence for
216 improvements of at least one surrogate endpoint. The remaining 13 (31.7%) indications were
217 supported by (very) strong statistical evidence for OS.

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

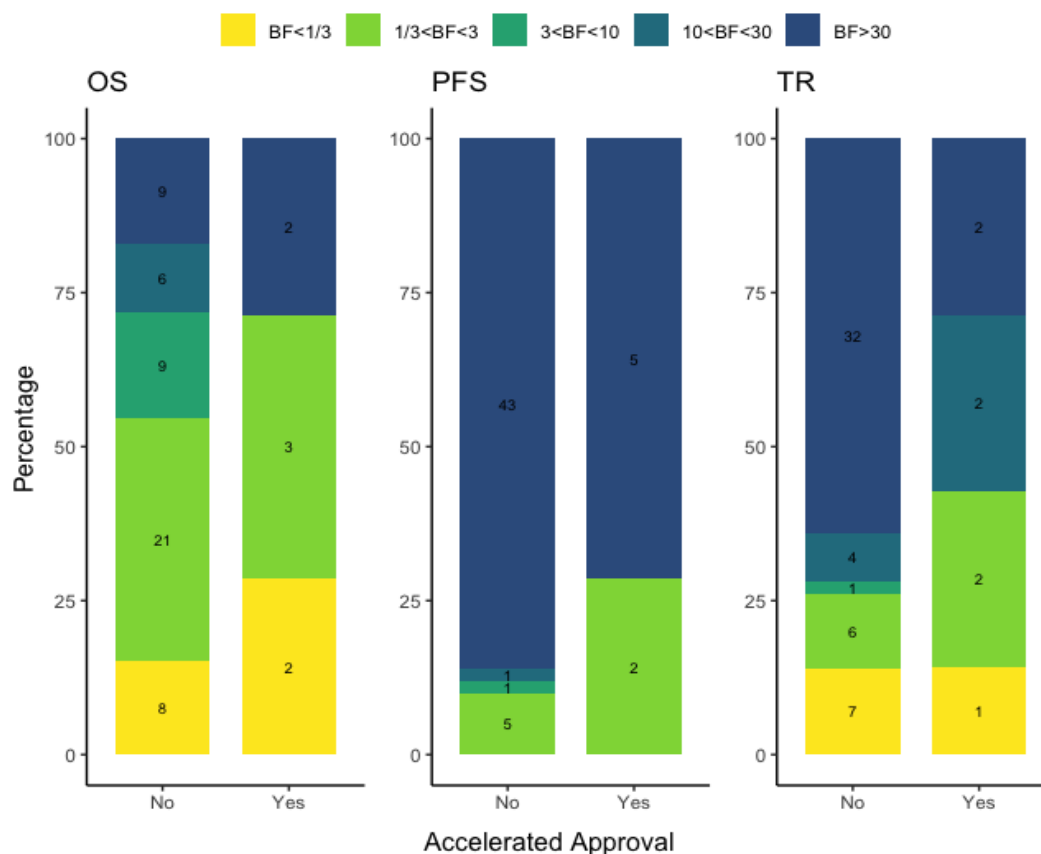
222 **Accelerated approval**

223 Strength of statistical evidence was consistently lower for accelerated approvals ($n = 7$)
224 across endpoints (Figure 4). No indications with accelerated approval lacked or provided (very)
225 strong statistical evidence for improvements across all three endpoints. For indications that received
226 non-accelerated approval, 5 of 58 indications lacked strong evidence for improvements on any
227 endpoint, while 12 (out of 58) had strong or very strong statistical evidence for improvements across
228 all three endpoints.

229 Five out of seven (71.4%) accelerated approval decisions were based on pro-null or
230 ambiguous statistical evidence for OS improvements, compared to 29 out of 53 (54.7%) non-
231 accelerated approval decisions. Two (20.0%) accelerated approvals were based on pro-null evidence
232 for OS improvements (active control: nivolumab; placebo control: panitumumab).

233 For PFS, no trial provided pro-null statistical evidence. The proportion of ambiguous
234 statistical evidence was higher for accelerated approvals ($2/7=28.6\%$) compared to non-accelerated
235 approval ($5/50=10.0\%$). For TR, the proportion of pro-null evidence was similar across approval
236 pathways (accelerated: $1/7=14.3\%$; non-accelerated: $7/50=14.0\%$), but accelerated approvals were

237 more frequently (2/7=28.6%) based on ambiguous statistical evidence compared to non-accelerated
 238 approvals (6/50=12.0%).



239
 240 Figure 4. Bar plot illustrating the proportion of trials supported by pro-null evidence ($BF < 1/3$),
 241 ambiguous evidence ($1/3 < BF < 3$), moderate pro-alternative evidence ($3 < BF < 10$), strong pro-alternative
 242 evidence ($10 < BF < 30$) and very strong pro-alternative evidence per endpoint and approval pathway.
 243 Note that indications based on multiple trials are not included in this figure

244 **Line of treatment**

245 For OS and PFS, strength of statistical evidence did not differ qualitatively between lines of
 246 treatment (see Table 1). Although median strength of statistical evidence was greater for first-line
 247 treatment than for second-line treatment, IQRs mostly overlapped. For TR, strength of statistical
 248 evidence was lowest for trials supporting drugs approved for third or later line of treatment.

249 **Cancer type**

250 There was no difference in strength of statistical evidence between solid and haematological
 251 cancer types (see Table 1).

252 **Approvals based on two RCTs**

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

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: $BF < \frac{1}{3}$; Light green: $\frac{1}{3} < BF < 3$, Dark green: $3 < BF < 10$, Turquoise: $10 < BF < 30$; Dark blue: $BF > 30$*

Drug	Outcome	Strength of Statistical Evidence (BF)		
		RCT 1	RCT 2	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	-

*Please note that for this indication the participant groups differed between the RCTs; RCT1 was conducted in previously treated patients and RCT2 in previously untreated patients.

254 Two drugs (28.6%) were approved despite both trials providing pro-null or ambiguous
255 statistical evidence. Of these, fulvestrant was approved based on two RCTs indicating that the
256 treatment did not perform better than the active control on OS (meta $BF = 0.07$) and TR (meta $BF =$
257 0.12). The other, ruxolitinib, was approved based on two RCTs providing ambiguous statistical
258 evidence for OS (meta $BF = 0.89$) and ambiguous statistical evidence for PFS from one trial ($BF =$
259 0.53)². For the other five drugs, one RCT with ambiguous statistical evidence was supplemented by
260 another RCT with at least moderate statistical evidence in favour of a treatment effect.

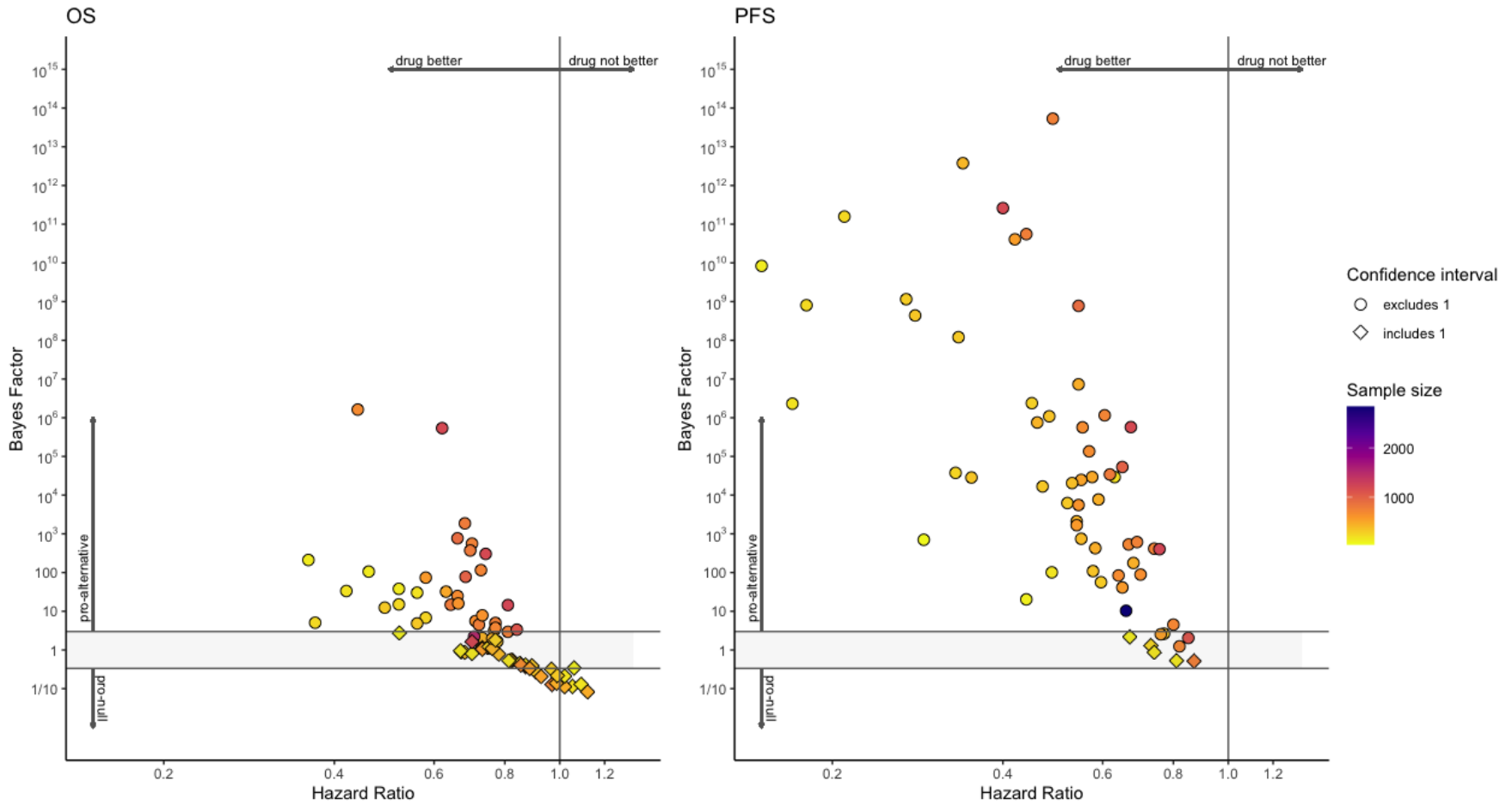
261 **Exploratory analysis**

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

263 In most cases, BF and 95% confidence intervals (CIs) were in agreement that an effect was
264 pro-alternative ($n=89$) or ambiguous ($n=28$; i.e., 95% CI included the null). However, in 12 cases the
265 95% CI indicated uncertainty, while the BF was more informative, indicating absence of efficacy
266 (lower left and right quadrant in Figure 5). Additionally, there were 12 effects for which the HR and
267 corresponding confidence intervals indicated efficacy, whereas the BF indicated ambiguous
268 statistical evidence.

269 There was no relationship between strength of statistical evidence and sample size ($r = -$
270 0.03).

² Note that for ruxolitinib phosphate evidence was pooled across trials with active and inactive comparators



271

272 Figure 5. *BFs plotted against HR for the 68 indications with one RCT. Shapes indicate whether or not the confidence interval of the HR included.*

273

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

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...	<ul style="list-style-type: none">- poor prognosis (despite available treatments)- few or highly toxic available treatments- prior failed treatments- different administration (e.g., oral medication)
Tolerability	<ul style="list-style-type: none">- limited crossover between treatment arms- low discontinuation rates
Positive interpretation of limited statistical evidence	<ul style="list-style-type: none">- meaningful clinical differences (e.g., differences being statistically not significant but effect size deemed clinically meaningful)- convincing interim analyses (e.g., very small <i>p</i>-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., statistically non-significant superiority trial interpreted as successful non-inferiority trial)
Primary endpoints different from OS, PFS, or TR	<ul style="list-style-type: none">- 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	<ul style="list-style-type: none">- control group deemed unethical

276

277 **Strength of statistical strength for primary endpoints**

278 The median strength of statistical evidence was greater for OS and PFS when these
279 outcomes were considered primary, compared to when they were not, although the interquartile
280 IQRs mostly overlapped (see Table 4). Additional exploration regarding the differences in evidential
281 strength between outcomes depending on which outcome was considered primary are presented in
282 the supplement.

283

Table 4. Median strength of statistical evidence for the primary and non-primary endpoints.

Endpoint	Primary	N trials	$M_{participants}$ (SD)	M_{events} (exp; con)	median BF [IQR]
OS	yes	25	672 (328)	223; 178	14.4 [2.9; 303.2]
	no	47	422 (281)	65; 61	1.1 [0.4; 4.9]
PFS	yes	46	479 (417)	120; 123	24767.8 [109; 7287000]
	no	36	555 (399)	258; 193	745.7 [84.0; 1.2*10 ⁸]
TR	yes	12	285 (185)	52; 32	47.3 [2.0; 886.0]
	no	70	551 (425)	87; 42	45.0 [1.0; 69880]

Note: exp: experimental arm; con: control arm; M_{events} refers to the primary outcome. These numbers are based both on individual trial endorsements and endorsements based on two RCTs combined. These numbers also combine trials for different cancer types which limits interpretability, as we expect differences in follow-up time and outcome choice between cancers.

284

285

DISCUSSION

286

Quantifying the strength of statistical evidence for efficacy for pivotal RCTs supporting novel

287

cancer drugs approved by the FDA in the last two decades, we were able to provide explicit evidence

288

that strength of statistical evidence was substantially lower for OS, arguably the most important

289

outcome for cancer patients, compared to surrogate outcomes. Most indications (58.7%, 44/75)

290

were approved without clear statistical evidence for OS improvements. While most of these

291

indications (33/44) were supplemented by strong statistical evidence for improvements on at least

292

one surrogate endpoint (i.e., PFS/TR), uncertainties regarding OS improvements remain. The present

293

analysis using the *BF* was more informative than traditional measures of uncertainty such as the

294

confidence interval, because it allowed us to disambiguate between absence of evidence and

295

evidence of absence.

296

Strength of statistical evidence for efficacy differed between approval pathways but not

297

lines of treatment or cancer types. Although few indications were approved through accelerated

298 approval, our results suggest that weaker statistical evidence is accepted for accelerated approval
299 decisions. While the FDA accepts higher levels of uncertainty for accelerated approval decisions due
300 to the use of surrogate endpoints or intermediate clinical endpoints [25], we provide quantification
301 of how much uncertainty the FDA considers acceptable. This again highlights the importance of
302 timely post-approval studies to confirm efficacy. This need is also recognized in the Consolidated
303 Appropriations Act, 2023 (H.R. 2617), enabling the FDA to require post-approval studies to be
304 *“underway prior to granting accelerated approval”*[26]. It remains unclear how consistently the FDA
305 will react to post-approval trials failing to confirm clinical benefits. As of 2021, one third of the novel
306 cancer drugs receiving accelerated approval until 2020 but subsequently failing to improve primary
307 endpoints in their post-approval studies, remained on the drug’s labelling (i.e., approved under the
308 accelerated pathway) or were converted to regular approval [27]. Clear regulations and consistent
309 action in response to post-approval trials is still lacking [27].

310 We observed indications receiving non-accelerated approval with absence of statistical
311 evidence or even statistical evidence for the absence of efficacy based on other considerations. For
312 example, favourable benefit-harm assessments were justified by expected benefits such as
313 improved quality of life and safety profiles. However, reporting of quality-of-life-outcomes is
314 incomplete [7] and not systematic, and a good safety profile is irrelevant in the absence of efficacy³.
315 In some cases, efficacy was determined in a manner different to the pre-registered protocol, for
316 instance through eyeballing Kaplan-Meier survival curves or comparing percentages. We observed
317 one instance in which a trial that failed to demonstrate superiority was re-interpreted as a non-
318 inferiority trial. Switching between superiority and non-inferiority interpretations after results are
319 known is problematic [28,29]. It would be preferable to supplement approval decisions with
320 additional trials, which would lead to stronger statistical evidence.

³ Please note that we are specifically referring to instances in which there is statistical evidence for the absence of efficacy. We do not mean instances in which there is no evidence that the novel treatment is better than standard care (i.e., a pro-null or ambiguous *BF* for active controlled trials).

321 In this paper, we primarily focused on cases with weak statistical evidence. However, there
322 was significant variation in the strength of evidence, with *BFs* for PFS and TR often suggesting solid
323 evidence for treatment effects. One might wonder, especially when time is of the essence, as it is
324 with cancer drugs, whether these drugs could be approved on the basis of weaker evidence.
325 However, the discussion of how much evidence is needed to determine efficacy based on surrogate
326 endpoints goes beyond the scope of this descriptive project, especially since this question cannot be
327 separated from the question of how well (*BFs* of) surrogate outcomes predict (*BFs* of) OS. These
328 questions warrant further investigation in future work.

329 **Strength and Limitations**

330 We used a comprehensive database of pre-approval trials, examined multiple endpoints,
331 and used a Bayesian approach to gain novel insights into the strength of evidence that allows for
332 interpretations that are in line with clinical decision making and may provide a more intuitive
333 perspective on evidence. The study also has several limitations. First, we focused on the statistical
334 evidence of RCTs supporting approval decisions. This does not reflect the full complexity of the
335 approval process and the variety of sources of evidence (e.g., quality of life, fewer side effects, method
336 of administration etc.) that might be considered. Nevertheless, RCTs generally provide the strongest
337 available evidence, as other sources of evidence (such as single-trial arms) are difficult to interpret.
338 Second, interpretation of the *BF* depends on the control group. While we differentiate between
339 active and inactive-controlled trials we did not classify whether comparators met the standard of
340 care. As a result, some of our *BF* might be an overestimation of the strength of evidence for efficacy.
341 Third, strength of statistical evidence is only one component of benefit assessment, and other
342 factors also play a role, including the effect size and risk of bias. Fourth, we restricted our analyses to
343 default priors to ensure comparability of our results across drug groups. Informed priors might be
344 used in future analysis of individual drugs to integrate the available evidence into the statistical
345 analysis. Lastly, we only included trials conducted to support approval decisions and did not include,

346 for instance, post-approval studies. Therefore, our results indicate the strength of statistical
347 evidence available to support initial approval decisions and do not necessarily reflect the current
348 strength of statistical evidence.

349 **Conclusion**

350 Regulatory decision making could be improved by using *BFs* to distinguish between drugs
351 with good statistical evidence, drugs that lack statistical evidence, and drugs with statistical evidence
352 against efficacy. We found that across the board the level of evidence for beneficial effects on OS is
353 low. Average strength of statistical evidence on OS was moderate only if OS was considered the
354 primary endpoint. While this suggests that evidence is better if the endpoint is considered primary,
355 OS should be important regardless of whether it is a primary outcome or not. Some drugs were even
356 approved without supporting statistical evidence on either OS, PFS or TR. These cases require a
357 transparent and clear explanation. In many cases the statistical evidence is ambiguous, calling for
358 additional trials, before or after approval, to reduce this uncertainty.

359

360

List of Abbreviations

BF	Bayes Factor
FDA	Food and Drug Administration
OS	Overall survival
PFS	Progression-free survival
RCT	Randomized controlled trial
TR	Tumor response

361

362

363 **Declarations**

364 **Ethics approval.** This study involved publicly available trial-level data. No ethical approval was
365 needed.

366 **Consent for publication.** Not applicable.

367 **Availability of data and materials.** The data that support the findings of this study are available from
368 <https://ceit-cancer.org/> and the OSF framework (<https://osf.io/4uhz7>) DOI 10.17605/OSF.IO/QZ7XY.

369 **Competing interests.** The authors declare that they have no competing interests.

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

371 **Authors' contributions:** *M.-M.P.*: Conceptualization, Data curation, Formal analysis, Methodology,
372 Project administration, Resources, Visualization, and Writing - original draft. *M.L.*: Formal analysis,
373 Software, and Writing - review & editing. *Y.A.d.V.*: Conceptualization, Supervision, and Writing -
374 review & editing. *L.G.H.*: Conceptualization, Investigation, and Writing - review & editing. *A.M.S.*:
375 Investigation and Writing - review & editing. *R.R.M.*: Conceptualization, Supervision, and Writing -
376 review & editing. *D.v.R.*: Conceptualization, Funding acquisition, Supervision, and Writing - review &
377 editing.

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

380

381

Acknowledgements

382 We would like to thank Jonathan Kimmelman for his insightful and constructive comments during
383 the review process.

384

385

References

- 386 1 Tafuri G, Stolp P, Trotta F, *et al.* How do the EMA and FDA decide which anticancer drugs make it
387 to the market? A comparative qualitative study on decision makers' views. *Ann Oncol.*
388 2014;25:265–9.
- 389 2 Ladanie A, Schmitt AM, Speich B, *et al.* Clinical Trial Evidence Supporting US Food and Drug
390 Administration Approval of Novel Cancer Therapies Between 2000 and 2016. *JAMA Netw Open.*
391 2020;3:e2024406.
- 392 3 Gyawali B, Hey SP, Kesselheim AS. Assessment of the Clinical Benefit of Cancer Drugs Receiving
393 Accelerated Approval. *JAMA Intern Med.* 2019;179:906.
- 394 4 Salas-Vega S, Iliopoulos O, Mossialos E. Assessment of Overall Survival, Quality of Life, and
395 Safety Benefits Associated With New Cancer Medicines. *JAMA Oncol.* 2017;3:382–90.
- 396 5 Vokinger KN, Kesselheim AS. Characteristics of trials and regulatory pathways leading to US
397 approval of innovative vs. non-innovative oncology drugs. *Health Policy.* 2019;123:721–7.
- 398 6 Michaeli DT, Michaeli T. Overall Survival, Progression-Free Survival, and Tumor Response Benefit
399 Supporting Initial US Food and Drug Administration Approval and Indication Extension of New
400 Cancer Drugs, 2003-2021. *J Clin Oncol Off J Am Soc Clin Oncol.* 2022;JCO2200535.
- 401 7 Gloy V, Schmitt AM, Dublin P, *et al.* The Evidence Base of US Food and Drug Administration
402 Approvals of Novel Cancer Therapies from 2000 to 2020. *Int J Cancer.* 2023;152:2474–84.
- 403 8 U.S. Food and Drug Administration. Guidance for Industry: E9 Statistical Principles for Clinical
404 Trials. 1998. <https://www.fda.gov/media/71336/download>
- 405 9 U.S. Food and Drug Administration. Qualification Process for Drug Development Tools. 2020.
- 406 10 Kemp R, Prasad V. Surrogate endpoints in oncology: when are they acceptable for regulatory
407 and clinical decisions, and are they currently overused? *BMC Med.* 2017;15:1–7.
- 408 11 Pignatti F, Jonsson B, Blumenthal G, *et al.* Assessment of benefits and risks in development of
409 targeted therapies for cancer — The view of regulatory authorities. *Mol Oncol.* 2015;9:1034–41.
- 410 12 Prasad V, Kim C, Burotto M, *et al.* The Strength of Association Between Surrogate End Points and
411 Survival in Oncology: A Systematic Review of Trial-Level Meta-analyses. *JAMA Intern Med.*
412 2015;175:1389–98.
- 413 13 Gronau QF, Ly A, Wagenmakers E-J. Informed Bayesian T-Tests. *Am Stat.* 2019;1–14.
- 414 14 Rouder JN, Speckman PL, Sun D, *et al.* Bayesian t tests for accepting and rejecting the null
415 hypothesis. *Psychon Bull Rev.* 2009;16:225–37.
- 416 15 Jeffreys H. *Theory of probability.* Oxford: Oxford University Press 1961.
- 417 16 Ladanie A, Speich B, Naudet F, *et al.* The Comparative Effectiveness of Innovative Treatments for
418 Cancer (CEIT-Cancer) project: Rationale and design of the database and the collection of
419 evidence available at approval of novel drugs. *Trials.* 2018;19:1–13.

- 420 17 Ladanie A, Ewald H, Kasenda B, *et al.* How to use FDA drug approval documents for evidence
421 syntheses. *BMJ*. 2018;362:k2815.
- 422 18 R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R
423 Foundation for Statistical Computing 2021. <https://www.R-project.org/>
- 424 19 Linde M, van Ravenzwaaij D, Tendeiro JN. *baymedr: Computation of Bayes Factors for Common*
425 *Biomedical Designs*. 2022. <https://github.com/maxlinde/baymedr>
- 426 20 Morey RD, Rouder JN, Jamil T, *et al.* BayesFactor: Computation of Bayes Factors for Common
427 Designs. 2022. <https://CRAN.R-project.org/package=BayesFactor> (accessed 19 July 2022)
- 428 21 Jamil T, Ly A, Morey RD, *et al.* Default “Gunel and Dickey” Bayes factors for contingency tables.
429 *Behav Res Methods*. 2017;49:638–52.
- 430 22 Heck DW, Gronau QF, Wagenmakers E-J. metaBMA: Bayesian model averaging for random and
431 fixed effects meta-analysis. Retrieved Doi. 2017. [https://cran.r-](https://cran.r-project.org/web/packages/metaBMA/metaBMA.pdf)
432 [project.org/web/packages/metaBMA/metaBMA.pdf](https://cran.r-project.org/web/packages/metaBMA/metaBMA.pdf) (accessed 20 March 2021)
- 433 23 Gronau QF, Heck DW, Berkhout SW, *et al.* A Primer on Bayesian Model-Averaged Meta-Analysis.
434 *Adv Methods Pract Psychol Sci*. 2021;4:25152459211031256.
- 435 24 Lee MD, Wagenmakers E-J. *Bayesian cognitive modeling: A practical course*. Cambridge
436 university press 2014.
- 437 25 U.S. Food and Drug Administration. Accelerated Approval. FDA. 2023.
438 [https://www.fda.gov/patients/fast-track-breakthrough-therapy-accelerated-approval-priority-](https://www.fda.gov/patients/fast-track-breakthrough-therapy-accelerated-approval-priority-review/accelerated-approval)
439 [review/accelerated-approval](https://www.fda.gov/patients/fast-track-breakthrough-therapy-accelerated-approval-priority-review/accelerated-approval) (accessed 30 January 2024)
- 440 26 New FDA Reform Legislation: Congress Gifts a “FDORA” for the Holidays. 2023.
441 [https://www.ropesgray.com/en/newsroom/alerts/2023/01/new-fda-reform-legislation-](https://www.ropesgray.com/en/newsroom/alerts/2023/01/new-fda-reform-legislation-congress-gifts-a-fdora-for-the-holidays)
442 [congress-gifts-a-fdora-for-the-holidays](https://www.ropesgray.com/en/newsroom/alerts/2023/01/new-fda-reform-legislation-congress-gifts-a-fdora-for-the-holidays) (accessed 16 February 2023)
- 443 27 Gyawali B, Rome BN, Kesselheim AS. Regulatory and clinical consequences of negative
444 confirmatory trials of accelerated approval cancer drugs: retrospective observational study.
445 *BMJ*. 2021;374:n1959.
- 446 28 U.S. Food and Drug Administration, Center for Drug Evaluation and Research. Non-Inferiority
447 Clinical Trials to Establish Effectiveness Guidance for Industry. 2016.
448 <https://www.fda.gov/media/78504/download>
- 449 29 Committee for Proprietary Medicinal Products. Points to consider on switching between
450 superiority and non-inferiority. *Br J Clin Pharmacol*. 2001;52:223–8.

451