Larger effect sizes in nonrandomized studies are associated with higher rates of EMA licensing approval

Djulbegovic, Benjamin; Glasziou, Paul; Klocksieben, Farina A; Reljic, Tea; VanDenBergh, Magali; Mhaskar, Rahul; Ioannidis, John P A; Chalmers, Iain

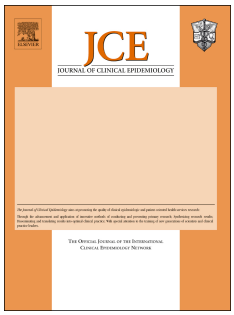
Published in:
Journal of Clinical Epidemiology

DOI:
10.1016/j.jclinepi.2018.01.011

Published: 01/06/2018

Document Version:
Peer reviewed version

Accepted Manuscript



Larger effect sizes in non-randomized studies are associated with higher rates of EMA licensing approval

Benjamin Djulbegovic, Paul Glasziou, Farina A. Klocksieben, Tea Reljic, Magali VanDenBergh, Rahul Mhaskar, John P.A. Ioannidis, Iain Chalmers

PII: S0895-4356(17)31070-3

DOI: 10.1016/j.jclinepi.2018.01.011

Reference: JCE 9596

To appear in: *Journal of Clinical Epidemiology*

Received Date: 21 September 2017

Revised Date: 17 December 2017

Accepted Date: 31 January 2018

Please cite this article as: Djulbegovic B, Glasziou P, Klocksieben FA, Reljic T, VanDenBergh M, Mhaskar R, Ioannidis JPA, Chalmers I, Larger effect sizes in non-randomized studies are associated with higher rates of EMA licensing approval, *Journal of Clinical Epidemiology* (2018), doi: 10.1016/j.jclinepi.2018.01.011.

Larger effect sizes in non-randomized studies are associated with higher rates of EMA licensing approval

Benjamin Djulbegovic[a,b] , Paul Glasziou[c], Farina A. Klocksieben[d], Tea Reljic[d], Magali VanDenBergh[e], Rahul Mhaskar[d], John P.A. Ioannidis[f] and Iain Chalmers[g]

[a] Department of Supportive Care Medicine, City of Hope, 1500 East Duarte Rd, Duarte, CA
[b] Department of Hematology, City of Hope, 1500 East Duarte Rd, Duarte, CA;
[c] Bond University, Gold Coast, 14 University Dr, Queensland, Australia
[d] Program for Comparative Effectiveness Research, University of South Florida, 12901 Bruce B Downs Blvd, Tampa, FL;
[e] H. Lee Moffitt Cancer Center & Research Institute, 12902 USF Magnolia Drive, Tampa, FL
[f] Stanford Prevention Research Center, Department of Medicine, and Department of Health Research and Policy, Stanford University School of Medicine; Department of Statistics, Stanford University School of Humanities and Sciences; and Meta-Research Innovation Center at Stanford, 450 Serra Mall, Stanford, CA,
[g] James Lind Initiative, Summertown Pavilion, Middle Way, Oxford OX2 7LG, UK

Word count:
Abstract: 187
Main text: 3,683

Corresponding author:

Benjamin Djulbegovic, MD, PhD

Departments of Supportive Care Medicine and Hematology
City of Hope
1500 East Duarte Rd.
Duarte, CA. 91010
+1 626 256 4673
E-mail: bdjulbegovic@coh.org

Abstract

Objectives: Evaluate how often the European Medicines Agency (EMA) has authorized drugs based on non-randomized studies and whether there is an association between treatment effects and EMA preference for further testing in RCTs.

Study Design and Setting: We reviewed all initial marketing authorizations in the EMA database on human medicines between 1995 and 2015 and included authorizations granted without randomized data. We extracted data on treatment effects and EMA preference for further testing in RCTs.

Results: Of 723 drugs, 51 were authorized based on non-randomized data. These 51 drugs were licensed for 71 indications. In the 51 drug-indication pairs with no preference for further RCT testing, effect estimates were large [OR 12.0 (95% CI: 8.1 to 17.9)] compared to effect estimates in the 20 drug-indication pairs for which future RCTs were preferred [OR 4.3 (95%CI 2.8 to 6.6)], with a significant difference between effects (p=0.0005).

Conclusions: Non-randomized data were used for 7% of EMA drug approvals. Larger effect sizes were associated with greater likelihood of approval based on non-randomized data alone. We did not find a clear treatment effect threshold for drug approval without RCT evidence.

Key words: dramatic effects- randomized trials- non-randomized studies-drug approval-quality of evidence-regulatory agencies

Running title: **Drug approval based on non-randomized evidence**

What is new?

- 7% of EMA drugs approved are based on data from non-randomized studies alone

- For authorizations that were granted on non-randomized data alone, larger estimated effect size is associated with a greater likelihood that authorization will be granted without EMA stating a preference for further testing in RCTs

- Depending on the theoretical framework used, between 2 to 4% of EMA approvals based on non-randomized data alone exhibited 'dramatic effects'

1. Introduction

Random allocation to treatment comparison groups is used to generate estimates of treatment effects free of distortion from allocation bias. Sufficiently large randomized clinical trials (RCTs) are typically required by drug regulatory agencies for deciding whether drugs should be licensed for use in clinical practice [1]. However, requiring RCTs is ethically questionable if there is insufficient uncertainty about the effects of treatments [2]. For example, uncertainty about effects of treatment is reduced when estimates of treatment effects are large ("dramatic effect"). Under these circumstances the effects of biases and the play of chance can be confidently ruled out without requiring testing in RCTs (e.g. insulin for treating diabetes or chest tube placement for treating pneumothorax).

What size of treatment effect is sufficiently dramatic to convince most people that the treatment differences observed are real and that the effects of bias and random error can be ruled out? Based on theoretical considerations, Glasziou and colleagues suggested that estimated risk ratios (RR) greater than 10 in comparison to no treatment or alternative treatments were needed to justify confident inferences about treatment effects based on non-randomized data [3]. The influential GRADE group also includes effect size as one of the criteria for upgrading the quality of observational evidence and suggests an RR of 5 as providing convincing evidence of an effect size[4, 5]. Retrospective analyses have indicated that between 0.05% and 2% of randomized trials have yielded effect sizes regarded as "dramatic" [6, 7]; however, they are rarely seen and validated in large trials [8].

A pragmatic way of assessing *how* large an estimated treatment effect must be to be regarded as *sufficiently* "dramatic" is to analyze the decisions of drug licensing authorities. The regulatory agencies have accepted that large treatment effects can sometimes obviate the need for data from RCTs. In 2012, the US Food and Drug Administration (FDA) introduced the "Breakthrough Therapy Designation" [9]; and in March and August 2016 the European Medicines Agency (EMA) launched the PRIME (Priority Medicines) and Adaptive Pathways programs to support approval of drugs demonstrating substantial improvement over existing therapies. In 2016, Hatswell and colleagues [10] reported the first systematic attempt to identify drugs approved without evidence from RCTs. They reviewed all drugs approved by the EMA and the FDA between January 1999 and May 2014, but did not include data on their effect sizes. In a small study examining FDA decisions to license nine cancer drugs (of which 6 were based on non-randomized comparisons) that had received Breakthrough Therapy Designation Kern concluded that the FDA would approve drugs under Breakthrough Therapy Designation if there was a doubling of treatment benefit compared to historical control groups [11].

Here, we report an analysis of all the authorized drugs listed in the database of the EMA, from its inception on January 26, 1995 to December 8, 2015 [www.ema.europa.eu]. We studied the nature of the comparison groups and the size of estimated treatment effects in drugs granted licenses based on data derived from non-randomized comparisons. We hypothesized that the estimated effect sizes of drugs for which EMA had not required evidence from RCTs would be larger than those for which EMA mentioned its request for subsequent confirmation in RCTs of estimates of effects.

We also wanted to assess whether there was any evidence of a threshold effect size beyond which testing in RCTs appeared to be deemed unnecessary.

2. Methods

2.1 Study selection

On December 8, 2015 we downloaded all information on medicines for humans from the EMA website which the agency received since its inception on January 26, 1995. Initial marketing authorization documents were reviewed by two investigators (FK, TR). All applications referring to EMA initial marketing authorizations for one or more indications based on non-randomized data were deemed eligible for inclusion (see Figure 1) [12].

2.2 Data Extraction

Data were extracted independently by two investigators (FK, TR) using a standardized form. Any disagreements were resolved by a third researcher (RM). The lead author (BD) verified a 20% sample of the data extractions. No major discrepancy in data extraction was found. Data were collected on study designs, disease characteristics, interventions, comparators (as described in section 2.3), and primary outcomes.

We extracted data on single events (outcomes) per patient from the EMA reports. In four instances, the EMA had calculated treatment effects based on repeated health outcomes measured in the same patients enrolled in the eligible studies. For example, drugs for treating hemophilia were typically tested in the same patients to assess the effects on repeated outcomes (bleeds). This violated the independence assumption underlying the statistical analyses, but probably biased our analyses against the experimental treatments, as people with repeated events could be expected to have

poorer responses to treatments. All analyses were based on the aggregate data reported in the EMA reports and none were based on individual-participant data.

2.3 Selection of comparators

When explicit descriptions of comparators were provided [e.g. events of interest (n) over the number of patients in experimental and control groups (N)], we used them in our analysis. In some cases, the EMA's authorization was based on predicted outcomes, as if the drug had been tested against placebo, no therapy, or obsolete standard treatment. For example, tyrosine kinase inhibitors developed after imatinib were not (originally) tested against imatinib but were compared against the previous standard treatments used to approve imatinib.

2.4 Effect size estimation

We used primary outcomes (as defined by the EMA) to calculate effect sizes (i.e. treatment effects). In some cases, only a proportion was provided as an effect estimate for the control arm (e.g. 15% response rate with standard treatment), without referencing total number of patients (N). Here we used the denominator (N) from the experimental arm as the denominator (N) in the control arm.

Where the EMA documents reported a minimum efficacy threshold for calculating a sample size, we used it to derive a hypothetical control, like that described above. For example, the EMA document refers to a FDA guideline stating that the statistical demonstration of a serious infection rate per person-year of less than 0.5 to 1.0 provides evidence of efficacy of the use of immunoglobulins for primary immunodeficiency disease (see Table S1).

Where the EMA documents neither provided nor cited data on the effects in comparators, we used terms from a patients-intervention-comparator-outcome (PICO) framework to search the literature and attempted to match experimental arms with control arms [13].

Sometimes the effects of the comparators were larger than in experimental arms, resulting in effect sizes favoring comparators (see Table S1). In these instances, EMA approvals appear to have been driven by considering secondary outcomes, or what was believed to be compelling biological rationale. For example, hematological remission was the primary outcome for comparing imatinib (and other similar tyrosine kinase inhibitors) with standard chemotherapy for chronic myeloid leukemia in blast crisis. Even though standard chemotherapy was superior to imatinib in terms of remission rates, the EMA authorized imatinib for this indication because it could induce cytogenetic and molecular remission (secondary outcomes), which almost never occurs with standard chemotherapy. Such secondary outcomes were used for 10 drugs: nine for assessment of the effect of tyrosine kinase inhibitors in chronic myeloid leukemia; and one for evaluating the effect of recombinant factor XIII for treatment of congenital FXIII deficiency.

2.5 Preference for subsequent RCT

We recorded any EMA mention of whether subsequent RCTs would be needed or desirable (even if this seemed unlikely to happen).

2.6 Study Appraisal

We appraised all included non-randomized studies using the Down and Black quality assessment instrument [14] after comparing it with the more recently developed

Cochrane tool [15] and finding that they generated similar results. Our assessment was based on publications when these were available, supplemented with information from the EMA documents.

2.7 Statistical Analysis

We used the Shapiro-Wilk test to assess whether the treatment effect sizes were normally distributed. When they were, we used t-tests to assess differences in effect sizes in drugs for which the EMA required further testing in RCTs and others. Otherwise we used non-parametric Kruskal-Wallis (K-W) tests. We also meta-analyzed data to compare the average effect size for the group of studies for which the EMA had indicated a wish for further testing in RCTs with other studies. Within each subgroup (RCTs requested, RCTs not requested), we summarized data under random-effects and tested for the differences in effect sizes between the two subgroups. Odds ratio (OR) was our metric of choice for the main analysis because, when control group success rates are already modestly high, RRs cannot reach very large values whereas OR values are unbounded to infinity. Since the most-widely used definition of 'dramatic' effects in the literature was based on RR [3, 4, 16], we calculated RR as well as absolute risk differences.

Because signal detection theory (Weber-Fechner law) suggests that people's perception of a difference is a function of the ratio of the signals (here the treatment versus control responses), we hypothesized that fewer approvals would mention the need for subsequent RCTs as treatment effect [i.e. ln(OR)] increased [17, 18]. To assess this postulated relationship, we used logistic regression, where the effect size expressed as the ln(OR) was used to predict the probability of not requiring an RCT.

To determine whether EMA decisions have reflected 'dramatic effects', our analysis used three definitions of the latter: (i) empirically-derived definition of 'dramatic effects' based on the results of this analysis (equal to OR ≥12- see section 3.5); (ii) an effect size with RR ≥ 5 (the GRADE criterion [4, 16]); or (iii) a RR ≥ 10 (as proposed by Glasziou et. Al [3]). We assessed how often drugs were authorized without requiring further RCTs under each of these criteria.

Finally, we performed sensitivity analyses to assess the robustness of our findings according to type of comparator (active vs. no active treatment), category of disease (cancer vs. chronic diseases vs. rare diseases vs. other), primary outcome (disease-oriented outcome, such as response rate vs. patient-oriented outcome, such as survival), and unit of analysis (patient or event). All analyses were performed using STATA, version 14 [19]. RevMan software was used to generate forest plots [20].

3. Results

3.1 Study selection and data sources

We reviewed all 4,109 initial marketing authorizations related to 3,351 unique active substances. Figure 1 shows the data selection process and reasons for exclusion. A major reason for exclusion was availability of RCT data (985 applications). Additionally, five applications were authorized based on a published case series only, rendering it impossible to ascertain if only "positive" outcomes had been reported. We excluded these authorizations because it was not possible to calculate effect sizes.

Overall, we identified 51 medicines that were authorized for 71 indications without evidence from RCTs (10 medicines were authorized for multiple indications). Table S1 presents the data included in our analyses. Briefly, published manuscripts were

available for 58 drug/indication pairs, conference abstracts for 2, and only the EMA initial marketing authorization documents for the remaining 11.

3.2 Use of non-randomized studies for authorizations

In all, 723 newly developed drugs were granted marketing authorization (672 drugs were authorized based on RCTs and 51 on non-randomized studies). Thus, 7% [51/723] of drugs were authorized based on non-randomized data.

3.3 Characteristics of included studies

Of the 71 drug-indication pairs authorized based on non-randomized data, 58% (41/71) were for treating cancer, particularly leukemias and lymphomas (Table S1). Another 27% (19/71) were for rare diseases, 8% (6/71) for chronic diseases, and 7% (5/71) for other health problems. The health problems leading to drug approvals were relatively rare, occurring in <1% of the general population.

Data on 42% (30/71) of comparators were extracted directly from the initial marketing authorization documents, 13% (9/71) from published manuscripts, and 45% (32/71) from a literature search using PICO. Comparators consisted of active treatment in 76% (54/71) of studies and no active treatment in 24% (17/71) of studies. Pre-defined response criteria (e.g. overall response, cytogenetic response, objective response, etc.) was the most commonly used primary outcome in 61% (43/71) of studies. Survival was the primary outcome in 6% (4/71) of studies. Two studies used continuous data for the primary outcome while the remainder used dichotomous data.The median number of patients in experimental arms was 80 (range 12 to 1710; the latter consisted of four single-arm studies combined). Data from prospective studies were used in 96% (68/71) of approvals and from retrospective studies in 4% (3/71). 76% (54/71) were single-arm

studies with no mention of controls; 7% (5/71) were single-arm studies with historic controls; and 17% (12/71) were multi-arm non-comparative studies.

3.4 Effect size and preference for subsequent RCT

Following drug-indication authorization, the EMA expressed a desire for subsequent RCT data in 28% (20/71) of cases. There was no statistically significant difference between number of patients in the experimental arms of studies for which the EMA wished to see treatment effects confirmed in RCTs compared with those for which no such desire was expressed (median: 78 vs.113; p=0.37 by K-W test).

The distribution of effect sizes ranged from OR of 1.06 to 2563 (Figure S1). Using raw data, the effect size was larger among drugs authorized without requiring confirmation of effects in RCTs [mean ln(OR): 2.88 +/- 1.70; (OR: 17.81 +/- 5.47)] compared with those for which the EMA would have liked confirmation in RCTs [mean ln(OR): 1.92 +/- 1.5 (OR: 6.82 +/- 4.48)], with a significant difference between effects (p=0.028) (Figure 2a). Similarly, using meta-analysis, the effect size was larger among drugs with no preference for RCT data [OR: 12.02 (95% CI: 8.08 to 17.89)] versus those that EMA stated a preference for an RCT [OR: 4.29 (95% CI: 2.80 to 6.58)]; with a significant difference between effects (p=0.0005) (Figure 2b). Figure S2 displays the subgroup analysis for all included indications.

3.5 Agreement with definition of dramatic effects

Figure 3 shows a flow-chart depicting the EMA preference for subsequent RCTs, as empirically determined from the data set, and based on the previously proposed criteria of “dramatic effects”. ORs ≥12 were observed in 31/51 (60.7%) of the indications for which the EMA did not require subsequent RCTs, compared with 6/20 (30%) (p=0.0195)

for which subsequent RCTs were deemed necessary. RRs ≥5 were noted in 20/50 (40%) indications for which EMA did not mention a preference for RCTs vs. 5/20 (25%) for which they did. RRs ≥10 were seen in 16/50 (32%) of studies for which the EMA did not desire a subsequent RCT as compared to 4/20 (20%) for which they did desire a subsequent RCT. Thus, the probability that the EMA will approve a drug based on 'dramatic effects' without requiring further RCTs was (i) 4.3% (31/723), (ii) 2.8% (20/723), and (iii) 2.2% (16/723) according to the three different 'dramatic effect' criteria. Figure 4 shows the cumulative distribution of the effect size as defined by OR, RR and absolute risk difference. As shown in Figure 4, ORs ≥12 were seen in 52% of the studies (4a); RRs ≥5 in 36%, and RRs ≥10 in 29% (4b). As also shown in Figure 4c, the absolute risk difference ranged from near 0% to almost 100%, with almost all values within that range having an equal probability of occurring.

We observed a statistically significant association between the log of effect size [ln (OR)] and the increase in the proportion of drugs authorized without requirement for evidence from RCTs [OR: 1.45; 95% CI: 1.02 to 2.07; p=0.037] (Figure S3). We also observed significantly larger effect sizes when the comparators were not active treatments [median: ln (OR): 3.78 (OR: 44.1) vs. ln (OR): 2.33 (OR: 10.3); p=0.038 by K-W test)]. No statistically significant associations were observed in the analysis of other subgroups [type of outcome (disease vs. patient-oriented), unit of analysis (event vs. patient), or type of disease category].

3.6 Quality assessment

Table S2 shows our quality assessments. The external validity for all studies included in our analysis was very low (judged to be zero out of maximum score 3). In addition, the

quality of studies which the EMA would have preferred to be followed by RCTs was lower for the reporting dimension than for studies where such consideration was not apparent from the EMA documents [median: 7 (range: 4 to 11) vs. 9 (range: 6 to 11); p=0.041 by K-W test]. Although the drugs were licensed for use in specified conditions, in none of the cases were we able to discern how patients had been selected from the source population, the proportion of patients who agreed to participate in the studies, and whether the studied patients were representative of patients seen in typical practice.

4. Discussion

We found that about 7% of drug approvals by the EMA have been granted without evidence from RCTs. Only 2 to 4% (depending on the definition of dramatic effect) of the drug approvals had shown dramatic effects in non-randomized data. On average, effect sizes were larger among studies for which the EMA did not require further testing in RCTs, an observation that is in accordance with the Weber-Fechner law (Figure S3) [17, 18]. However, we found no clear evidence of a specific ‘dramatic effect’ threshold (Figure 3).

Our findings are in line with the only other empirical study on this topic: after examining FDA approvals of 9 drugs (6 based on non-RCT comparison) for 10 indications, Kern (2016) concluded that medications had been approved under Breakthrough Therapy Designation if they resulted in a doubling of improvement compared with outcomes observed using historical controls [11].

Several other considerations apart from the presence of dramatic effects seem likely to be operating in these licensing decisions, but the EMA does not provide an explicit rationale for approving new drugs without comparison studies.In principle, the trustworthiness of EMA decisions could be assessed by seeing whether estimates of the effects of drugs authorized based on non-randomized studies is supported by estimates derived from subsequent RCTs. We found only one such example (an RCT of imatinib confirmed the effect seen in a non-randomized phase II trial in the treatment of chronic myeloid leukemia) [21]. Several other drugs (for example, ofatumumab for refractory chronic lymphocytic leukemia, and tyrosine kinase inhibitors for treatment of chronic myeloid leukemia) were subsequently tested in RCTs, but not against the comparator that had been used in the non-randomized studies used to support authorization [22].

In the absence of empirical evidence, EMA decision-making needs to be judged against accepted best methodological practices. For example, the International Conference on Harmonization E10 states: (E 10, 2001) [23]

> "The inability to control bias restricts use of the external control design to situations in which the effect of treatment is **dramatic** and the usual course of the disease highly predictable. In addition, use of external controls should be limited to cases in which the endpoints are objective and the impact of baseline and treatment variables on the endpoints is well characterized**."**

For some of the conditions listed in Table S1, these requirements do not seem to have been fulfilled. Without having data with clear delineation of numerators and denominators (n/N) in experimental and control arms, it is impossible to estimate

treatment effects reliably. In only 24% (17/71) of indications did the EMA provide data on numerators and denominators for control arms, relying instead on implicit comparisons. For example:

> "Given the non-comparator design of the study, efficacy outcomes cannot be compared with response in a placebo or active arm. Furthermore, comparison of efficacy results with published data is considered to be inappropriate as there does not appear to be any published data from comparable studies. …Consequently, **efficacy has been assessed on the basis of response rates in comparison to what would be expected by expert clinical evaluation and by comparison with previous experience in this type of patient"** (our emphasis).

Without evidence, how can experts know what they claim to know? [24-26] The EMA documents often indicated that a current treatment was not considered effective, but reference to systematic reviews of existing evidence, formal surveys of experts, and analyses of registry data were rare [27].

Given the imputations that we were forced to use in estimating treatment effects, especially for control groups, some of our treatment effect estimates are inevitably approximations. That is, in our attempts to translate the EMA judgments into the effect sizes, we frequently understood that the EMA assumed very low (often equal to zero) event rates such as response rate or survival in the control arm (Table S1). This seems likely to be reason that we observed some improbably high effect sizes. The quality of the studies considered by the EMA and critical assessment of surrogate outcomes [28] and other features was unsatisfactory (Table S2) raising further questions about

reliability of the EMA judgments. The situation can be improved if the regulatory agencies such as the EMA and the FDA adopt a more formal system, such as GRADE, in their evaluation of approval based on non-RCTs. To upgrade the quality of observational evidence at the level of RCTs, GRADE takes into account not only the effect size but also pays attention to dose-response gradient and requests explanation for plausible residual confounding.[4, 5]

It is important to note, however, that the lack of rigor in choosing comparators results is a problem for further drug development. Once a drug has been authorized by regulatory agencies it tends to be regarded as the appropriate comparator for testing new drugs, reinforcing the preceding unsatisfactory comparison [22, 29].

5. Conclusion

Licensing drugs for use in clinical practice must strike a balance between failing to approve effective drugs and approving ineffective or dangerous drugs [30]. Because values and risk tolerance differ among people, a perfect technical solution to this challenge is unlikely. Even so, more explicit guidelines regarding the circumstances in which regulatory agencies are likely to approve drugs based on non-randomized data would help investigators, practitioners and public to take more informed decisions about assessing the effects of drugs. Regulatory agencies such as the EMA could facilitate this process through greater transparency about the basis for choosing comparators [28, 31-33]; providing references to the systematic reviews of historical comparisons that should underpin its methods; and being more explicit about the basis for its choices of decision thresholds.

Acknowledgments

As no authors work for the pharmaceutical industry, we are grateful to Professor Patrick Vallance, GSK, for helpful insights related to our findings and for promoting funding for this project under the project title “When Randomized Clinical Trials (RCTs) are not necessary? Quantifying dramatic intervention effects” (PI: Djulbegovic). However, GSK made no restrictions, conditions or any requirements whatsoever related to the design, analysis or report of our study. No GSK product was included in our analysis.

We are also grateful to Professor Douglas Altman, Centre for Statistics in Medicine and UK EQUATOR Centre, Nuffield Department of Orthopaedics, Oxford, UK; Professor Michael Rawlins, School of Hygiene and Tropical Medicine, London, UK; Professor Silvio Garattini, Mario Negri Institute for Pharmacological Research, Milan, Italy; and Dr. Gianluca Baio, Department of Statistical Science, University College, London UK for their useful feedback and comments on an earlier version of the manuscript

Conflict of interest

None.

Authorship

The conception and design of the study were completed by BD, PG, JI, and IC. Acquisition of data was done by BD, FAK, TR, MV, and RM. Analysis of the data was done by BD, FAK, and TR. Interpretation of data was handled by BD, PG, FAK, TR, RM, JI, and IC. Drafting the manuscript was completed by BD and FAK, with critical revisions done by all authors. All authors gave final approval of the completed manuscript.

Funding

GlaxoSmithKline provided funding for this study (award no: 3000029429). The funder had no role in the protocol development, study execution, or decision to publish.

Legends

Figure 1 PRISMA diagram of study selection process.

Figure 2. Distribution of the effect sizes as a function of the EMA's preference to have results confirmed in RCTs. The effect size was significantly larger among drugs authorized without confirmation in RCTs compared with those for which the EMA would have liked confirmation in further RCTs; a) raw data b) meta-analytic aggregate

Figure 3. The EMA preference for RCTs according to three different criteria of 'dramatic effects': a) empirically determined; b) an effect size with RR (risk ratio) ≥ 5 (the GRADE criterion[4, 16]); or c) a RR ≥ 10 (as proposed by Glasziou et al.[3]). (Denominator of n=70 in calculation of effect size is because in one study primary outcome was based on continuous data, thus preventing calculation of RR)

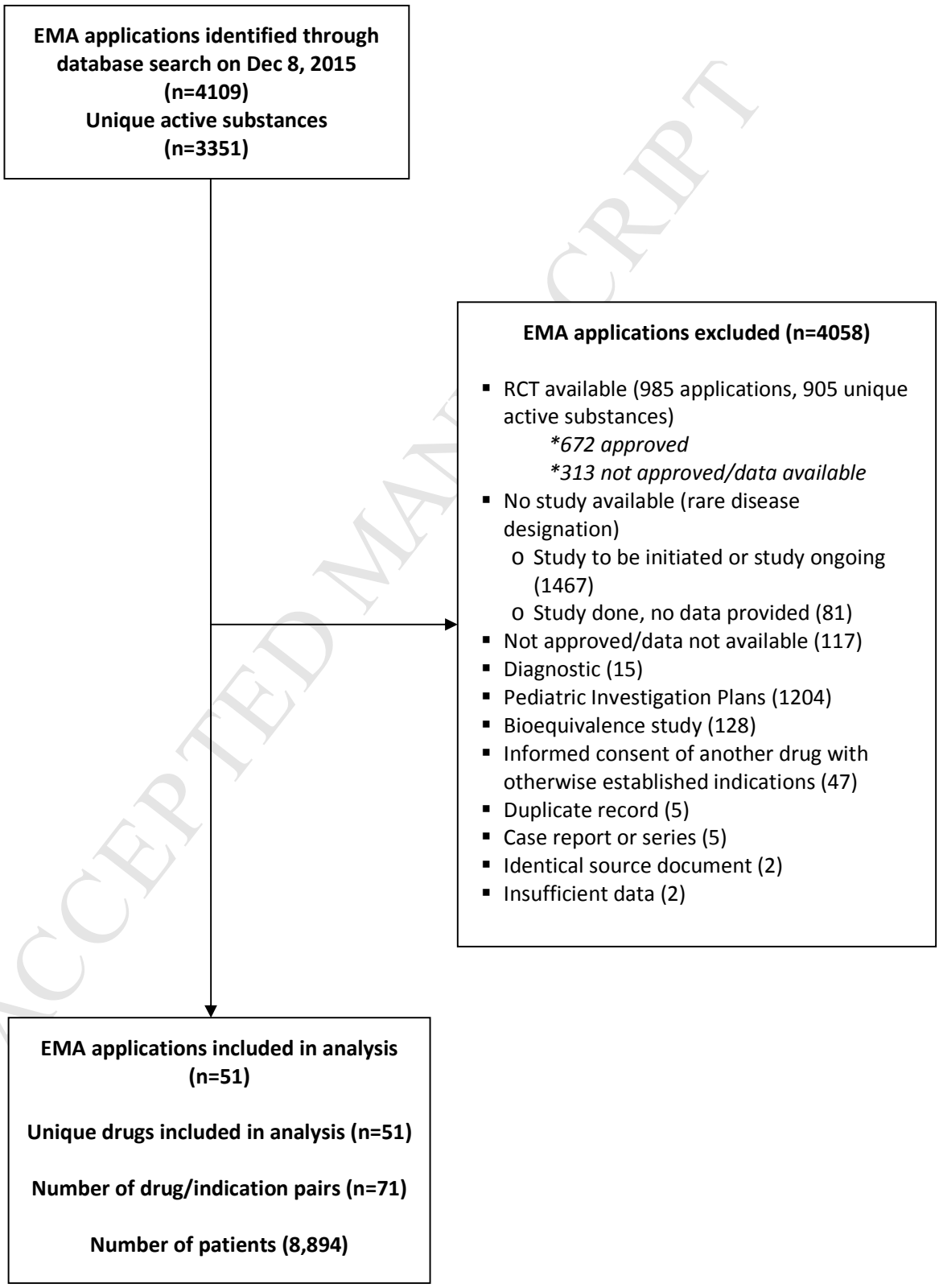
Figure 4. Cumulative distribution of the effect size as defined by OR, RR (risk ratio) and absolute risk difference. "Dramatic effects" were defined as: a) empirically derived[ORs ≥12, which was seen in 52% of the studies]; b) according to GRADE criterion [RRs ≥5 seen in 36% of studies], and according to Glasziou and colleagues [3][RRs ≥10 seen in 29% studies] (see also Fig 3).

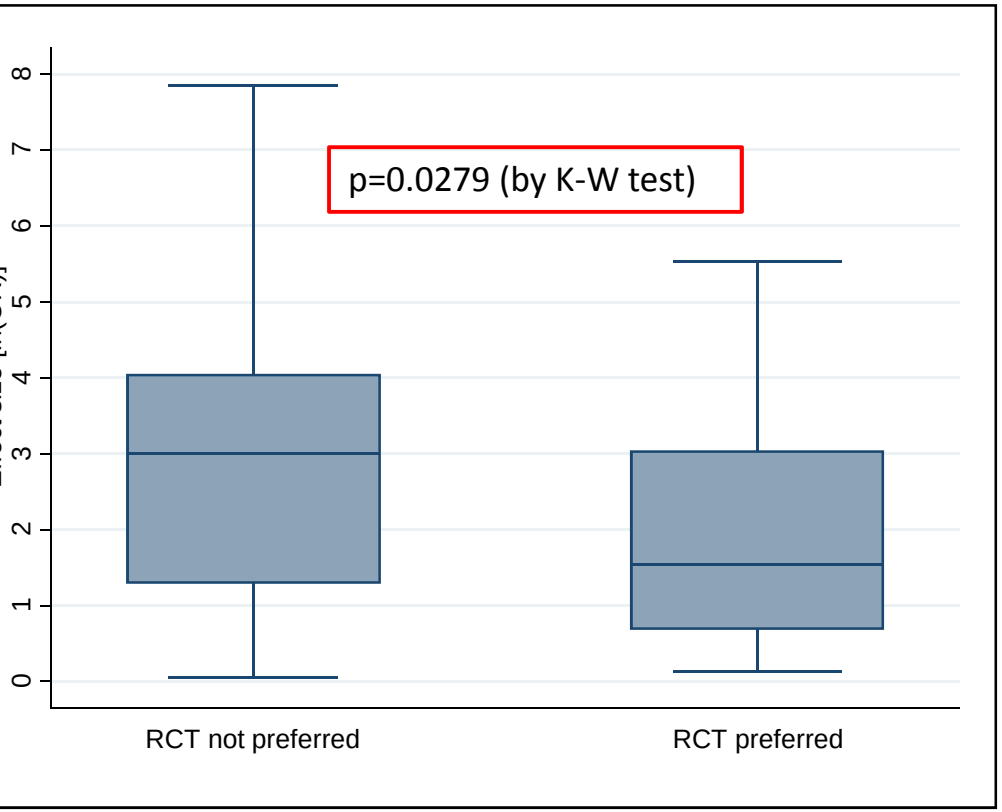
References

1. U.S. Department of Health and Human Services Food and Drug Administration: Guidence for industry: providing clinical evidence of effectiveness for human drug and biological products. In. Washington,D.C.: Center for Drug Evaluation and Research; 1998.
2. Djulbegovic B: Articulating and responding to uncertainties in clinical research. *J Med Philosophy* 2007, 32:79-98.
3. Glasziou P, Chalmers I, Rawlins M, McCulloch P: When are randomised trials unnecessary? Picking signal from noise. *BMJ* 2007, 334(7589):349-351.
4. Brozek JL, Akl EA, Compalati E, Kreis J, Terracciano L, Fiocchi A, Ueffing E, Andrews J, Alonso-Coello P, Meerpohl JJ *et al*: Grading quality of evidence and strength of recommendations in clinical practice guidelines part 3 of 3. The GRADE approach to developing recommendations. *Allergy* 2011, 66(5):588-595.
5. Guyatt GH, Oxman AD, Montori V, Vist G, Kunz R, Brozek J, Alonso-Coello P, Djulbegovic B, Atkins D, Falck-Ytter Y *et al*: GRADE guidelines: 5. Rating the quality of evidence--publication bias. *J Clin Epidemiol* 2011, 64(12):1277-1282.
6. Miladinovic B, Kumar A, Mhaskar R, Djulbegovic B: Benchmarks for detecting 'breakthroughs' in clinical trials: empirical assessment of the probability of large treatment effects using kernel density estimation. *BMJ Open* 2014, 4(10):e005249.
7. Nagendran M, Pereira TV, Kiew G, Altman DG, Maruthappu M, Ioannidis JP, McCulloch P: Very large treatment effects in randomised trials as an empirical marker to indicate whether subsequent trials are necessary: meta-epidemiological assessment. *BMJ* 2016, 355:i5432.
8. Pereira Tv HRIIJA: Empirical evaluation of very large treatment effects of medical interventions. *JAMA* 2012, 308(16):1676-1684.
9. Sherman RE, Li J, Shapley S, Robb M, Woodcock J: Expediting Drug Development — The FDA's New “Breakthrough Therapy” Designation. *New England Journal of Medicine* 2013, 369(20):1877-1880.
10. Hatswell AJ, Baio G, Berlin JA, Irs A, Freemantle N: Regulatory approval of pharmaceuticals without a randomised controlled study: analysis of EMA and FDA approvals 1999-2014. *BMJ Open* 2016, 6(6):e011666.
11. Kern KA: Trial Design and Efficacy Thresholds for Granting Breakthrough Therapy Designation in Oncology. *J Oncol Pract* 2016, 12(8):e810-817.
12. Liberati A, Altman DG, Tetzlaff J, Mulrow C, GÃ¸tzsche PC, Ioannidis JPA, Clarke M, Devereaux PJ, Kleijnen J, Moher D: The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *PLoS Med* 2009, 6(7):e1000100.
13. Straus S, Richardson W, Glasziou P, Haynes R: Evidence-base Medicine. How to practice and teach EBM.: Churchill Livingstone; 2005.
14. Downs S, Black N: The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. *J Epidemiol Community Health* 1998, 52:377 - 384.

15. Sterne JA, Hernan MA, Reeves BC, Savovic J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I *et al*: ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. *BMJ* 2016, 355:i4919.
16. Guyatt GH, Oxman AD, Sultan S, Glasziou P, Akl EA, Alonso-Coello P, Atkins D, Kunz R, Brozek J, Montori V *et al*: GRADE guidelines: 9. Rating up the quality of evidence. *J Clin Epidemiol* 2011, 64(12):1311-1316.
17. Lanzara RG: Weber's law modeled by the mathematical description of a beam balance. *Math Biosci* 1994, 122(1):89-94.
18. Stevens SS: Neural events and the psychophysical law. *Science* 1970, 170(962):1043-1050.
19. STATA C: STATA, ver. 14. In*.* College Station, TX; 2013.
20. Review Manager (RevMan). In*.*, 5.3 edn. Copenhagen: The Nordic Cochrane Centre, The Cochrane Collaboration; 2014.
21. O'Brien SG, Guilhot F, Larson RA, Gathmann I, Baccarani M, Cervantes F, Cornelissen JJ, Fischer T, Hochhaus A, Hughes T *et al*: Imatinib Compared with Interferon and Low-Dose Cytarabine for Newly Diagnosed Chronic-Phase Chronic Myeloid Leukemia. *New England Journal of Medicine* 2003, 348(11):994-1004.
22. Naci H, Wouters OJ, Gupta R, Ioannidis JPA: Timing and Characteristics of Cumulative Evidence Available on Novel Therapeutic Agents Receiving Food and Drug Administration Accelerated Approval. *Milbank Q* 2017, 95(2):261-290.
23. E 10: Guidance for Industry Choice of Control Group and Related Issues in Clinical Trials. In*.* Edited by U.S. Department of Health and Human Services Food and Drug Administration. Washington DC; 2001.
24. Silverman WA: Gnosis and random allotment. *Controlled Clin Trials* 1981, 2:161-164.
25. Silverman WA, Chalmers I: Casting and drawing lots: a time-honoured way of dealing with uncertainty and for ensuring fairness. *BMJ* 2001, 323:1467-1468.
26. Chalmers I: Invalid health information is potentially lethal. *BMJ* 2001, 322(7292):998.
27. Behera M, Kumar A, Soares HP, Sokol L, Djulbegovic B: Evidence-based medicine for rare diseases: implications for data interpretation and clinical trial design. *Cancer Control* 2007, 14(2):160-166.
28. Mann H, Djulbegovic B: Comparator bias: why comparisons must address genuine uncertainties. *J R Soc Med* 2013, 106(1):30-33.
29. Salanti G, Dias S, Welton NJ, Ades AE, Golfinopoulos V, Kyrgiou M, Mauri D, Ioannidis JP: Evaluating novel agent effects in multiple-treatments meta-regression. *Stat Med* 2010, 29(23):2369-2383.
30. Djulbegovic B, Hozo I: When Should Potentially False Research Findings Be Considered Acceptable? *PLoS Medicine* 2007, 4(2):e26.
31. Garattini S, Chalmers I: Patients and the public deserve big changes in evaluation of drugs. *BMJ* 2009, 338:b1025.
32. Banzi R, Bertele V, Garattini S: EMA's transparency seems to be opaque. *The Lancet* 2014, 384(9957):1847.
33. Banzi R, Gerardi C, Bertele V, Garattini S: Conditional approval of medicines by the EMA. *BMJ* 2017, 357:j2062.

Figure 1. PRISMA diagram





b)

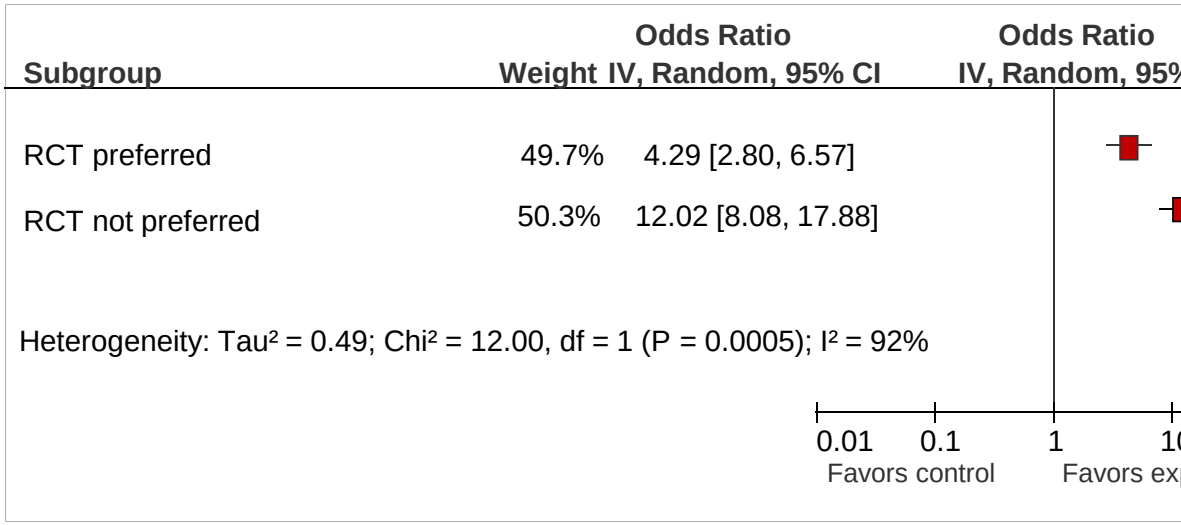
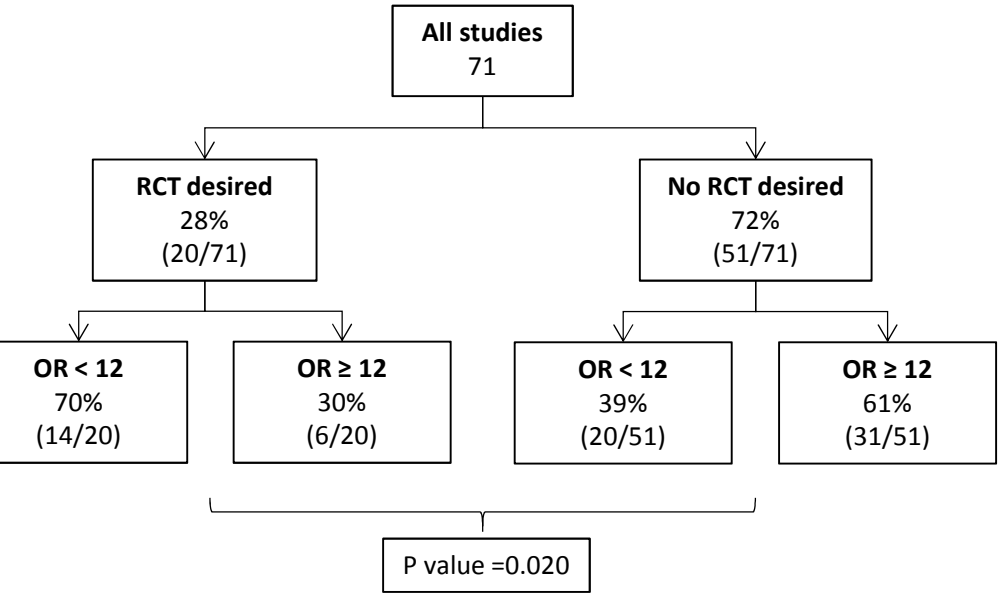
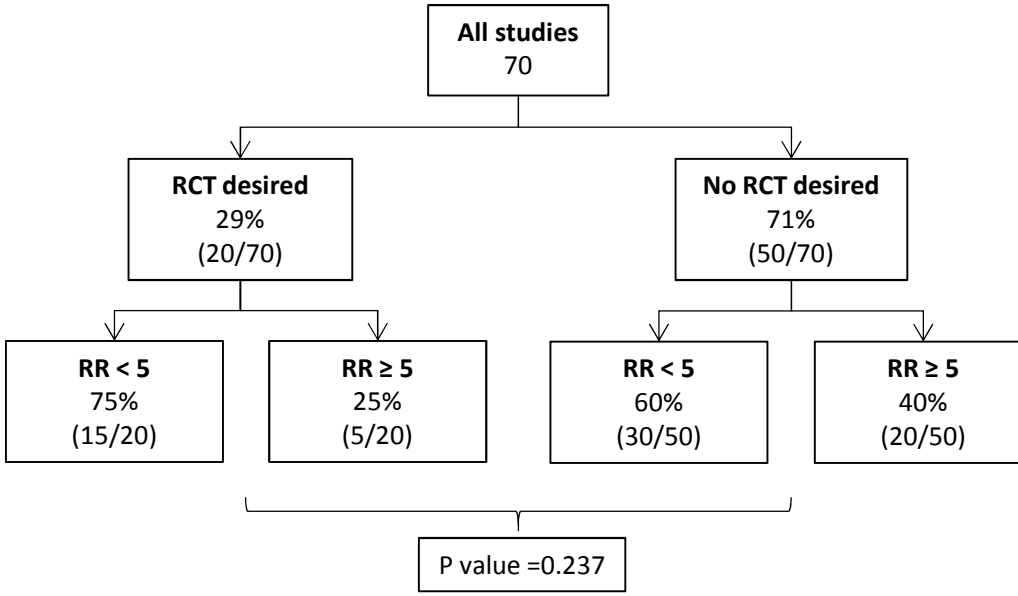


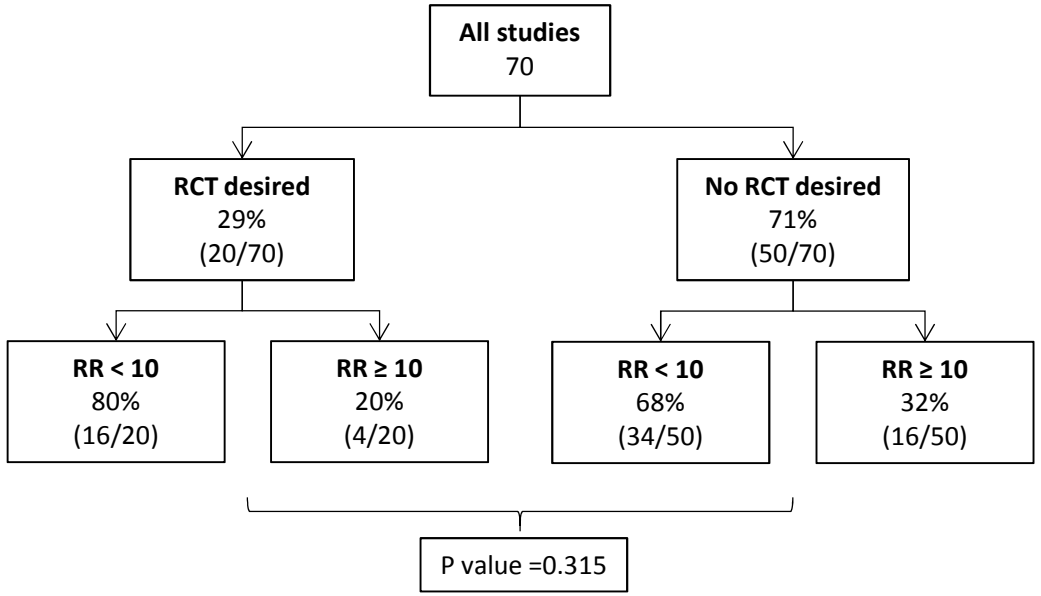
Figure 2 Distribution of the effect sizes as a function of preference by EMA to have results confirmed in RCTs a) raw d
b) meta-analytic aggregate

b)

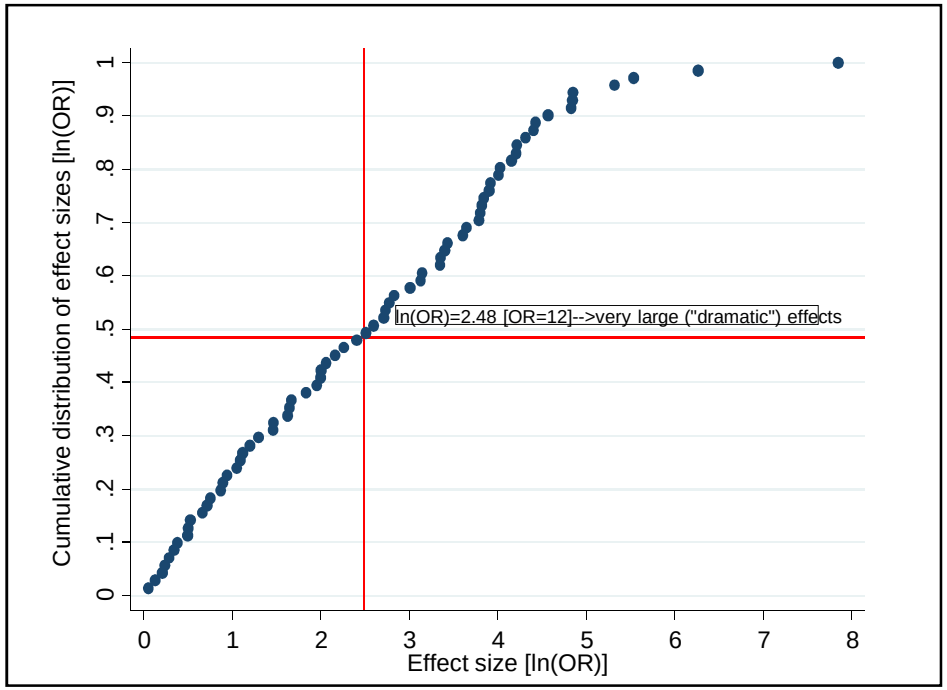

All studies
71
RCT desired
28%
(20/71)
No RCT desired
72%
(51/71)
OR < 12
70%
(14/20)
OR ≥ 12
30%
(6/20)
OR < 12
39%
(20/51)
OR ≥ 12
61%
(31/51)
P value =0.020


All studies
70
RCT desired
29%
(20/70)
No RCT desired
71%
(50/70)
RR < 5
75%
(15/20)
RR ≥ 5
25%
(5/20)
RR < 5
60%
(30/50)
RR ≥ 5
40%
(20/50)
P value =0.237

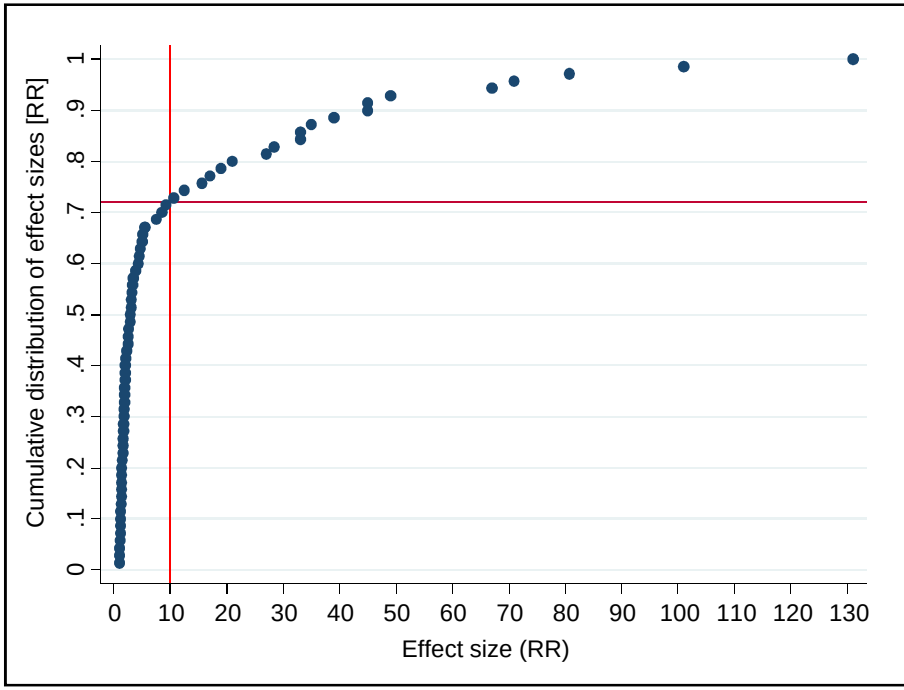
c)


All studies
70
RCT desired
29%
(20/70)
No RCT desired
71%
(50/70)
RR < 10
80%
(16/20)
RR ≥ 10
20%
(4/20)
RR < 10
68%
(34/50)
RR ≥ 10
32%
(16/50)
P value =0.315

gure 3



b)



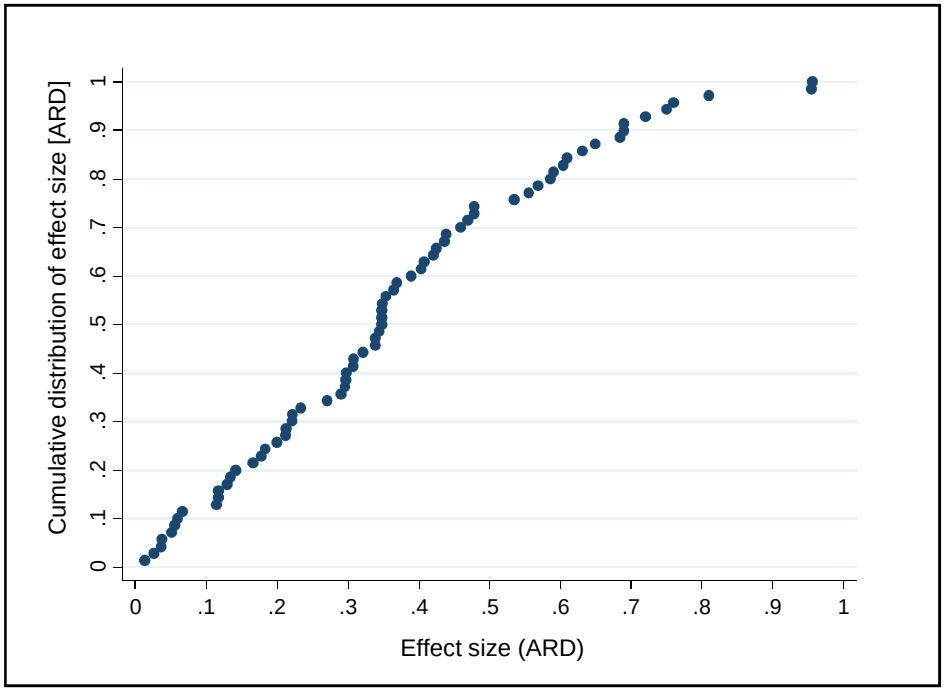


Figure 4 Cumulative distribution of the effect s
as defined by a) odds ratio, b) risk ratio and c)
absolute risk difference.