

ORIGINAL ARTICLE



Benchmarking single-arm studies against historical controls from non-small cell lung cancer trials – an empirical analysis of bias

Kelli Snyders^a, Doah Cho^b, Jun Hee Hong^a, Sally Lord^{b,c}, Rebecca Asher^b, Ian Marschner^{b,d} and Chee Khoon Lee^{a,b}

^aCancer Care Centre, St George Hospital, Kogarah, Australia; ^bNational Health and Medical Research Council Clinical Trials Centre, Camperdown, Australia; ^cSchool of Medicine, The University of Notre Dame, Darlinghurst, Australia; ^dDepartment of Statistics, Macquarie University, Sydney, Australia

ABSTRACT

Background: Recent trials of novel agents in 'rare' molecular subtypes of non-small cell lung cancer (NSCLC) have used single-arm trial designs and benchmarked outcomes against historical controls. We assessed the consistency of historical control outcomes using docetaxel data from published NSCLC randomized controlled trials (RCTs).

Material and methods: Advanced NSCLC RCTs including a docetaxel monotherapy arm were included. Heterogeneity in tumor objective response rates (ORRs), progression-free survival (PFS) and overall survival (OS), and correlations between outcomes and year of trial commencement were assessed.

Results: Among 63 trials ($N = 10,633$) conducted between 2000 and 2017, ORR ranged from 0% to 26% ($I^2 = 76.1\%$, $p_{\text{heterogeneity}} < .0001$). Mean of the median PFS was 3.0 months (range: 1.4–6.4), 3-month PFS ranged from 25% to 85% ($I^2 = 86.0\%$, $p_{\text{heterogeneity}} < .0001$). Mean of the median OS was 9.1 months (range: 4.7–22.9), 9-month OS ranged from 23% to 79% ($I^2 = 83.0\%$, $p_{\text{heterogeneity}} < .0001$). Each later year of trial commencement was associated with 0.3% ($p = .046$), 0.5% ($p = .11$) and 0.9% ($p = .001$) improvement in ORR, 3-month PFS and 9-month OS rates, respectively.

Conclusions: There was significant heterogeneity and an improving trend in docetaxel outcomes across trials conducted over 20 years. Benchmarking biomarker-targeted agents against historical controls may not be a valid approach to replace RCTs. Innovative study designs involving a concurrent control arm should be considered.

Abbreviations: NSCLC: non-small cell lung cancer; RCTs: randomized controlled trials; ORR: objective response rate; PFS: progression-free survival; OS: overall survival; CIs: confidence intervals; RECIST: Response Evaluation Criteria In Solid Tumors; PFS3: 3-month progression-free survival rate; OS9: 9-month overall survival rate

ARTICLE HISTORY

Received 24 July 2019

Accepted 25 September 2019

Introduction

Advanced non-small cell lung cancer (NSCLC) is a highly heterogeneous disease and is being classified into biologically distinct 'rare' subgroups according to molecular biomarkers [1]. Novel therapies are being developed to target these biomarkers and prominent examples, such as osimertinib [2] for the *EGFR* mutant subtype, have shown strong activity from early phase single-arm trials. While the value of randomized trials with concurrent controls is widely recognized and traditionally have been required for drug approval and reimbursement, there have been calls to remove this requirement for novel targeted therapies when early phase results are compelling [3–5]. Supportive arguments include threatened equipoise, fewer patients available for trial recruitment and delayed access to active therapies.

Tumor objective response rate (ORR) is commonly selected as the primary efficacy endpoint in single-arm trials of biomarker-targeted therapies. In the absence of a

concurrent control arm, the ORR results are benchmarked against published results for standard-of-care therapy, in many cases chemotherapy, in similar patient groups ('historical' controls). The validity of this approach for targeted therapies has not been adequately evaluated to justify its use. In order for a historical control to provide a valid comparison, a minimal requirement is consistent outcomes across past trials. Differences in treatment outcomes between trials of the same therapies can be used as a measure of the potential for selection bias using this approach. To test this hypothesis, we examined ORR for docetaxel reported in past randomized controlled trials (RCTs). Docetaxel was chosen because it has been a commonly recommended second-line therapy option by international clinical practice guidelines [6–9] and frequently used as a control treatment in second-line RCTs in advanced NSCLC conducted over the past two decades. Furthermore, it has been used as a historical control benchmark for newer agents in contemporary trials [10–12].

CONTACT Doah Cho  doah.cho@ctc.usyd.edu.au  NHMRC Clinical Trials Centre, The University of Sydney, Locked Bag 77, Camperdown, NSW 1450, Australia

 Supplemental data for this article can be accessed [here](#).

© 2019 Acta Oncologica Foundation

Material and methods

Search strategy

We searched the MEDLINE and EMBASE databases and hand-searched conference proceedings to identify eligible RCTs published between January 2000 and December 2017. RCTs were eligible if they assessed second- and subsequent-line therapies in advanced NSCLC, included docetaxel monotherapy as a treatment arm, and reported ORR ([Supplementary Search Strategy](#)).

Data extraction and analysis

From each trial, we extracted patient demographic characteristics, patient recruitment period and ORR (defined as complete and partial tumor responses). Using the published Kaplan–Meier's curves, we estimated 3-month progression-free survival (PFS) and 9-month overall survival (OS) rates. We selected 3-month PFS and 9-month OS rates as outcome variables as they are clinically relevant time points in this patient population. We computed the 95% confidence intervals (CIs) using the Exact method. Where it was not possible to calculate the CI using the normal approximation, exact binomial CIs were constructed for the proportions. The inverse-variance weighting used for the meta-analysis was calculated based on these CIs [13]. We assessed heterogeneity in treatment effects across trials using the I^2 statistic which quantifies the percentage of variability between estimates beyond sampling error, and tested the null hypothesis of no difference using the Cochrane Q test. We computed pooled overall estimates of ORR, 3-month PFS and 9-month OS rates using the fixed-effects inverse-variance weighted method. We assessed changes in demographic characteristics over time using weighted linear regression and examined variation in treatment outcomes according to patient recruitment period using meta-regression. All analyses were performed using Stata (StataCorp. 2017; Stata Statistical Software: Release 15, StataCorp LLC, College Station, TX, USA).

The searches, data extraction and risk of bias assessments using the Cochrane risk of bias tool [14] were performed independently by K.S. and J.H.H. Differences were resolved by consensus.

Results

Trial and patient characteristics

We identified 63 eligible RCTs comprising 10,633 patients recruited between 1994 and 2016 ([Supplementary Figure S1](#), [Supplementary Reference list](#)). Docetaxel 75 mg/m² was administered on day 1 every three weeks in 53 (84%, $N = 9270$) trials originating outside Japan, while docetaxel 60 mg/m² was administered on day 1 every three weeks in 10 (16%, $N = 1363$) trials originating in Japan. Fifty-eight trials reported performance status and 37 trials included patients with performance status 2. The majority of patients had good performance status (0 or 1: 92%, 2: 5%) and were

male (67%). Docetaxel was predominately used in the second-line setting (68%). RECIST criteria were used to assess response in 54 (86%) trials while two trials used Southwest Oncology Group criteria and seven trials used World Health Organization criteria ([Supplementary Table S1](#)). Molecular subtype was reported in only 10 (16%) trials.

Outcome data

ORR, PFS and OS

ORR ranged from 0% to 26% ($I^2 = 76.1\%$, $p_{\text{heterogeneity}} < .0001$) and the pooled estimate was 8% (95% CI 8–9%, [Figure 1](#)). The mean of the median PFS was 3.0 months (range 1.4–6.4) and the 3-month PFS rate ranged from 26% to 85% ($I^2 = 86.0\%$, $p_{\text{heterogeneity}} < .0001$, [Supplementary Figure S2](#)). The mean of the median OS was 9.1 months (range 4.7–22.9) and the 9-month OS rate ranged from 23% to 79% ($I^2 = 83.0\%$, $p_{\text{heterogeneity}} < .0001$, [Supplementary Figure S3](#)).

Trend over time

The proportion of females ($p = .03$) and proportion of patients with performance status 0 and 1 ($p < .0001$) included in trials increased over time ([Supplementary Table S1](#)). We observed a trend for docetaxel outcomes to improve over time. Each later year of trial commencement was associated with 0.3% ($p = .046$), 0.5% ($p = .11$) and 0.9% ($p = .001$) improvements in ORR, 3-month PFS and 9-month OS rates, respectively ([Figure 2](#), [Supplementary Table S2A](#), [Supplementary Figure S4](#)). Variations in ORR ($R^2 = 16\%$), 3-month PFS ($R^2 = 4\%$) and 9-month OS ($R^2 = 24\%$) rates across trials, as explained by year of trial commencement, were small ([Supplementary Table S2A](#)). The association between year of trial commencement and ORR, 3-month PFS and 9-month OS rates were not significant after accounting for other trial demographic variables ([Supplementary Table S2B](#)).

Sensitivity analyses

In sensitivity analyses, when we analyzed trials originating in Japan separately, we observed qualitatively similar results to the overall analysis ([Figure 1](#), [Supplemental Figure S2](#) and [Figure S3](#)). When we analyzed only those trials ($N = 54$, 86%) that utilized RECIST for tumor response assessment, the ORR ranged from 0% to 26% ($I^2 = 78\%$, $p_{\text{heterogeneity}} < .0001$). When we limited our analysis to only those trials of phase 3 design ($N = 33$, 52%), the ORR ranged from 0% to 20% ($I^2 = 82\%$, $p_{\text{heterogeneity}} < .0001$).

Risk of bias

The risk of bias was low for all 58 published trials, but unclear for the five unpublished trials.

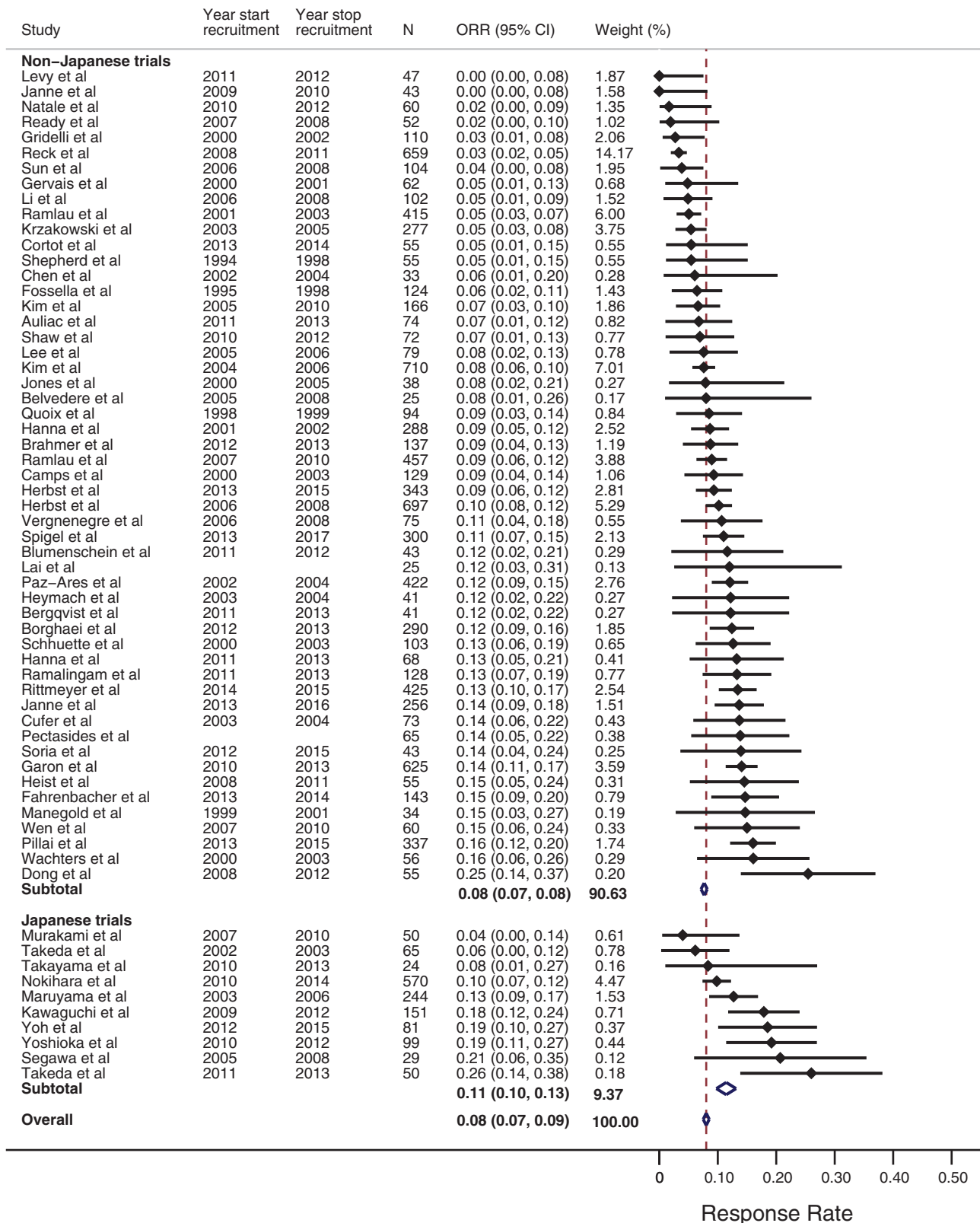


Figure 1. Forest plot of objective response rates for docetaxel. The point estimate of response rate for each trial is represented by the filled diamond, and the horizontal line crossing the diamond represents the 95% confidence interval (CI). The open diamond represents the pooled overall effect size. Japanese trials refer to studies originating in Japan. Non-Japanese trials may include Japanese subjects but refer to studies originating outside Japan.

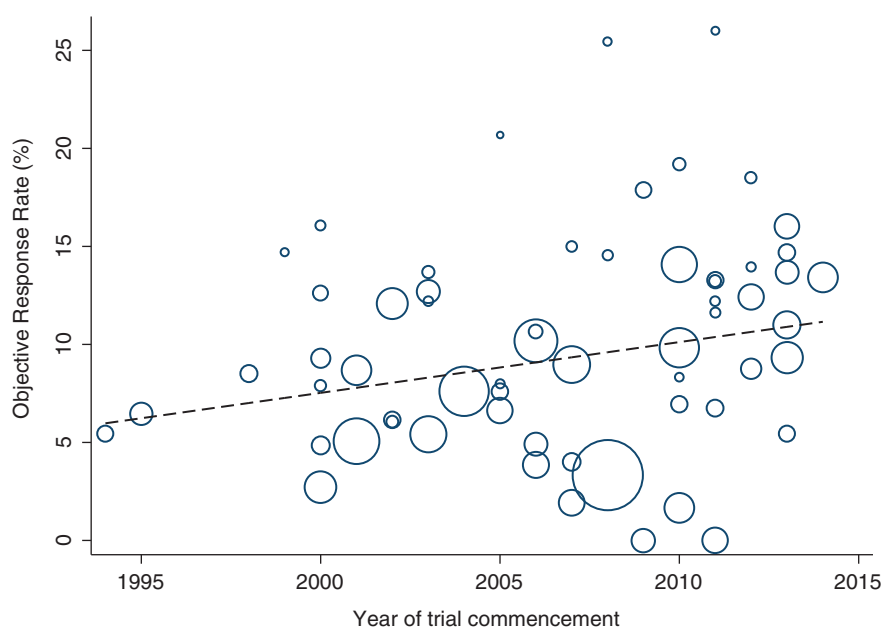


Figure 2. Distribution of tumor objective response rate over year of trial commencement. Each circle represents a trial. The circle size is inversely proportional to the standard error of the response rate. The dashed line is a fitted regression line of the relationship of objective response rate with year of trial commencement.

Discussion

Our finding of substantial variability in ORR (0–26%) for patients assigned to docetaxel monotherapy in advanced NSCLC RCTs demonstrates the fundamental weakness of adopting an approach of benchmarking single-arm studies with historical controls. We observed an improving trend for ORR, 3-month PFS and 9-month OS over time. However, only a small proportion of the overall heterogeneity of treatment outcomes was explained by the year of trial commencement. Furthermore, the association between year of trial commencement and treatment outcomes were not significant after accounting for other trial demographic variables. This indicates that data from a few contemporary trials will be insufficient to derive robust benchmarks for future single-arm studies.

Our study extends previous research by examining a larger number of trials in a more contemporary treatment period within a different tumor histology. In a cross-trial comparison of advanced ovarian cancer patients treated with the same platinum combination therapy, statistically significant differences in PFS and OS were reported [15]. In advanced breast cancer patients receiving cyclophosphamide, methotrexate and 5-fluorouracil, small but statistically significant differences in ORR and OS were found across different trials conducted in different eras even after adjustment for known baseline prognostic factors [16]. Differences in population characteristics cannot be completely accounted for with statistical modeling, as it is not possible to adjust for unknown or unmeasured prognostic factors. More importantly, there is no simple approach to refine historical benchmarks to account for baseline prognostic factors and eliminate selection bias. Arguably there is greater ambiguity in selecting a benchmark for molecularly targeted therapies as there is less prior prognostic data in this molecularly defined subset. In this example of docetaxel in advanced

NSCLC, only 10% of trials reported on molecular subtype and any clinical outcome differences on docetaxel between a molecularly defined versus undefined population could not be assessed.

The small improvements in ORR, 3-month PFS and 9-month OS rates with docetaxel over time are unlikely related to this chemotherapy. An increased proportion of patients with better prognostic characteristics may have contributed to this trend. In our study, the proportion of males, who have a shorter life expectancy than females [17], decreased over time and more patients with good performance status were recruited to later year studies. Other factors, including advances in imaging to identify patients with metastatic but low volume disease, increased availability of treatment options, better supportive care and improved management of treatment-related toxicities possibly resulting in longer treatment duration, may partly explain the improved treatment outcomes.

Our findings also caution against the suggested approach of using an absolute ORR threshold in single-arm trials as a definitive endpoint for drug approval. A recent review of single-agent anticancer drugs that received accelerated regulatory approval in the US suggests that high ORR exceeding 30% is associated with breakthrough activity [18]. However, in our docetaxel meta-analysis, we demonstrate that some trials of docetaxel monotherapy in NSCLC closely approached this threshold of ‘compelling evidence’.

The main strength of our study is that we have performed a comprehensive review of RCTs evaluating outcomes of docetaxel administered in similar settings in more than 10,000 patients. As these trials were conducted over almost two decades, it also allowed us to investigate trends in treatment outcomes. Several meta-analyses of second-line therapies in advanced NSCLC have previously been published to estimate the comparative effectiveness and safety of the various therapeutic options [19–22] and assessed for predictive

[20,21] and prognostic factors [22]. To the best of our knowledge, this is the first study focusing on heterogeneity in treatment outcome from the same docetaxel monotherapy as second-line therapy in advanced NSCLC.

There are also limitations in our study. We did not have access to individual patient data to allow comprehensive evaluation of baseline prognostic characteristics. Very few trials reported the molecular subtypes of patients. Tumor response assessment methods varied between trials; however, a sensitivity analysis performed using trials that utilized RECIST did not substantially alter the main finding of our study. Investigator-assessed versus centrally determined response status and variations in the timing and frequency of tumor assessments across trials may have also affected outcome assessments. Importantly, these factors are rarely accounted for when benchmarking results of single-arm studies against the outcomes of these historical controls.

Emerging challenges in designing, planning and recruiting to clinical trials of targeted therapies are not limited to advanced NSCLC, but extend to other cancers with molecularly defined 'rare' subgroups. A basket trial design with multi-arms to simultaneously examine various novel targeted agents with a concurrent control arm of patients treated with standard-of-care therapies, when available, has been proposed [23] and will overcome the challenges associated with benchmarking against a historical comparator. Improvement of such a design by including multi-institution collaboration and an adaptive design has also been recommended. Furthermore, novel, more sensitive markers of treatment effect, such as cell-free circulating DNA, incorporated into more traditional efficacy outcomes, such as tumor response, to form a composite endpoint may help address the problem of less reliable treatment estimates from smaller studies.

In conclusion, our meta-analysis of docetaxel monotherapy for the second line treatment of advanced NSCLC demonstrates that benchmarking against docetaxel as 'historical' controls may not be a valid approach to replace the use of a concurrent control arm in clinical trials in rare molecular subtypes of NSCLC. Claims of compelling improvement in treatment outcomes from single-arm studies cannot be easily quantified. Future trials with innovative designs involving a concurrent control arm should be considered.

Acknowledgements

We acknowledge the editorial support provided by Dr Sherilyn Goldstone, PhD (NHMRC Clinical Trials Centre).

Disclosure statement

The authors report no conflicts of interest.

References

- [1] Pikor LA, Ramnarine VR, Lam S, et al. Genetic alterations defining NSCLC subtypes and their therapeutic implications. *Lung Cancer*. 2013;82(2):179–189.
- [2] Jänne PA, Yang J-H, Kim D-W, et al. AZD9291 in EGFR inhibitor-resistant non-small-cell lung cancer. *N Engl J Med*. 2015;372(18):1689–1699.
- [3] Ng TL, Camidge DR. AURA 3: the last word on chemotherapy as a control arm in EGFR mutant NSCLC? *Ann Transl Med*. 2017;5(51):S14.
- [4] Selaru P, Tang Y, Huang B, et al. Sufficiency of single-arm studies to support registration of targeted agents in molecularly selected patients with cancer: lessons from the clinical development of crizotinib. *Clin Transl Sci*. 2016;9(2):63–73.
- [5] Simon R, Blumenthal GM, Rothenberg ML, et al. The role of non-randomized trials in the evaluation of oncology drugs. *Clin Pharmacol Ther*. 2015;97(5):502–507.
- [6] Azzoli CG, Baker S Jr, Temin S, et al. American Society of Clinical Oncology Clinical Practice Guideline update on chemotherapy for stage IV non-small-cell lung cancer. *J Clin Oncol*. 2009;27(36):6251–6266.
- [7] Felip E, Gridelli C, Baas P, et al. Metastatic non-small-cell lung cancer: consensus on pathology and molecular tests, first-line, second-line, and third-line therapy: 1st ESMO Consensus Conference in Lung Cancer; Lugano 2010. *Ann Oncol*. 2011;22(7):1507–1519.
- [8] Besse B, Adjei A, Baas P, et al. 2nd ESMO Consensus Conference on Lung Cancer: non-small-cell lung cancer first-line/second and further lines of treatment in advanced disease. *Ann Oncol*. 2014;25(8):1475–1484.
- [9] Masters GA, Temin S, Azzoli CG, et al. Systemic therapy for stage IV non-small-cell lung cancer: American Society of Clinical Oncology Clinical Practice Guideline Update. *J Clin Oncol*. 2015;33(30):3488–3515.
- [10] Kitagawa C, Iwasaku M, Kogure Y, et al. Phase II study of weekly amrubicin for refractory or relapsed non-small cell lung cancer. In *Vivo* (Athens, Greece). 2019;33(1):163–166.
- [11] Wu F, Zhang S, Xiong A, et al. A phase II clinical trial of apatinib in pretreated advanced non-squamous non-small-cell lung cancer. *Clin Lung Cancer*. 2018;19(6):e831–e842.
- [12] Lee JS, Lee KH, Cho EK, et al. Nivolumab in advanced non-small-cell lung cancer patients who failed prior platinum-based chemotherapy. *Lung Cancer*. 2018;122:234–242.
- [13] Sanchez-Meca J, Marin-Martinez F. Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychol Methods*. 2008;13(1):31–48.
- [14] Higgins JP, Altman DG, Gotzsche PC, et al. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*. 2011;343(2):d5928.
- [15] Markman M. The dangers of "cross-trial" and "cross-retrospective experience" comparisons: examples employing data in the peer-reviewed ovarian cancer literature. *Cancer*. 2007;109(10):1929–1932.
- [16] Lee CK, Lord SJ, Stockler MR, et al. Historical cross-trial comparisons for competing treatments in advanced breast cancer – an empirical analysis of bias. *Eur J Cancer* (Oxford, England: 1990). 2010;46(3):541–548.
- [17] World Health Organization. Global health observatory (GHO): female life expectancy [Internet]. United Nations. Geneva, Switzerland: World Health Association; c1948–2019; [cited 2019 Apr 19]. Available from: https://www.who.int/gho/women_and_health/mortality/situation_trends_life_expectancy/en/
- [18] Oxnard GR, Wilcox KH, Gonen M, et al. Response rate as a regulatory end point in single-arm studies of advanced solid tumors. *JAMA Oncol*. 2016;2(6):772–779.
- [19] Jin Y, Sun Y, Shi X, et al. Meta-analysis to assess the efficacy and toxicity of docetaxel-based doublet compared with docetaxel alone for patients with advanced NSCLC who failed first-line treatment. *Clin Ther*. 2014;36(12):1980–1990.
- [20] Crequit P, Chaimani A, Yavchitz A, et al. Comparative efficacy and safety of second-line treatments for advanced non-small cell lung cancer with wild-type or unknown status for epidermal growth factor receptor: a systematic review and network meta-analysis. *BMC Med*. 2017;15(1):193.

- [21] Vickers AD, Winfree KB, Cuyun Carter G, et al. Relative efficacy of interventions in the treatment of second-line non-small cell lung cancer: a systematic review and network meta-analysis. *BMC Cancer*. 2019;19(1):353.
- [22] Stroh M, Green M, Cha E, et al. Meta-analysis of published efficacy and safety data for docetaxel in second-line treatment of patients with advanced non-small-cell lung cancer. *Cancer Chemother Pharmacol*. 2016;77(3):485–494.
- [23] Lopez-Chavez A, Thomas A, Rajan A, et al. Molecular profiling and targeted therapy for advanced thoracic malignancies: a biomarker-derived, multiarm, multihistology phase II basket trial. *J Clin Oncol*. 2015;33(9):1000–1007.