Quantitative and qualitative assessment of real world data comparative effectiveness research of systemic therapies in lung oncology: A systematic review

Bas J.M. Peters, PhD PharmD, Vivi E.M.T. Janssen, BSc, Franz M. Schramel, MD PhD, Ewoudt M.W. van de Garde, PharmD

*Department of Clinical Pharmacy, St. Antonius Hospital, Utrecht/Nieuwegein, The Netherlands
*Department of Clinical Pharmacology and Pharmacy, VU University Medical Center, Amsterdam, The Netherlands
*Department of Pulmonary Diseases, St Antonius Hospital, Utrecht/Nieuwegein, The Netherlands
*Division of Pharmacoepidemiology and Clinical Pharmacology, Department of Pharmaceutical Sciences, Utrecht University, Utrecht, The Netherlands

ARTICLE INFO

Article history:
Received 15 April 2016
Accepted 11 July 2016
Available online xxx

Keywords:
Comparative effectiveness research
Observational research
Systematic review
Lung cancer
Lung oncology
Real world data

ABSTRACT

Introduction: The growing interest in comparative effectiveness research (CER) based on data from routine clinical practice also extends towards lung oncology. Although CER studies using real world data (RWD) have the potential to assist clinical decision-making, concerns about the quality and validity of studies with observational data subsist. The primary objective of the present study is to assess the current status of observational CER in the field of lung oncology, both quantitatively as qualitatively.

Methods: We performed a systematic electronic literature database search in MEDLINE and EMBASE (up to 1 July 2015). The quality of all selected studies was assessed according to the Good ReseArch for Comparative Effectiveness (GRACE) checklist.

Results: The first selection included 657 publications. After screening the corresponding abstracts and full-text papers, 38 studies remained. A total of 36 studies included patients with advanced NSCLC. The comparison of the effectiveness of gefitinib versus erlotinib was the main objective in 22% of the studies. The median number of patients per study was 202 (range 21–10064). The number of publications increased over the years whereas the quality score remained stable over the years with several common shortcomings (checklist items M5, D1, D4, D6).

Discussion: The growing interest in clinical oncology CER studies using RWD is reflected in an increasing number of publications in the recent years. The studies have several common methodological shortcomings possibly limiting their applicability in clinical decision-making. To fulfil the promise of RWD CER in lung oncology effort should be continued to overcome these shortcomings.

© 2016 Elsevier Ltd. All rights reserved.

Contents

1. Introduction .................................................................................................................. 6
2. Methods .......................................................................................................................... 6
   2.1. Literature search ........................................................................................................ 6
   2.2. Data extraction ........................................................................................................ 6
   2.3. Quality assessment ................................................................................................. 6
   2.4. Statistical analysis ................................................................................................. 6
3. Results ........................................................................................................................... 6
   3.1. Descriptive results .................................................................................................. 6
   3.2. Quality assessment ................................................................................................. 7
4. Discussion ...................................................................................................................... 11

* Corresponding author at: Department of Clinical Pharmacy, St. Antonius Hospital, Utrecht/Nieuwegein, The Netherlands.
E-mail address: b.peters@antoniusziekenhuis.nl (B.J.M. Peters).

http://dx.doi.org/10.1016/j.canep.2016.07.005
1877-7821/ © 2016 Elsevier Ltd. All rights reserved.
1. Introduction

The allocation of $1.1 billion by the American Recovery and Reinvestment acts of 2009 to further develop comparative effectiveness research (CER) as an alternative strategy to obtain relevant data for informed decisions in healthcare highlights the promise of and high demand for CER. The main argument for this investment was the increasing health expenditure driven by rapid development of new medical technologies, in many fields including (lung) oncology.

The main purpose of comparative effectiveness research (CER) is to assist consumers, clinicians, purchasers, and policy makers to make informed decisions that will improve health care at both the individual and population levels. The two key elements of CER are a direct comparison of interventions and studying them in patients who are typical of day-to-day clinical care [1,2].

Traditional randomized clinical trials (RCTs) basically never meet the latter two criteria because inclusion criteria for patients are considered not representative of the “real-world” population [3]. In contrast, observational research has the potential to be of major additional value by means of comparing interventions in a real world setting. Moreover, the use of real world data (RWD) has many other advantages such as: low costs, real-time data and a potentially larger number of patients and outcomes. Although it is nowadays well recognized that there is a need for RWD CER studies to assist clinical decision-making, concerns about the quality and validity of studies conducted with observational data persist. Limitations inherent to the use of observational data should therefore carefully be acknowledged and appropriately addressed. The objective of the present study is to assess the current status of observational CER on systemic therapy in lung oncology, both quantitatively and qualitatively. Of all types of cancer, lung cancer is the most common cause of cancer mortality in spite of many systemic treatment options.

2. Methods

2.1. Literature search

To obtain an overview of all CER studies in the field of lung oncology we conducted a systematic electronic literature database search in MEDLINE using PubMed and EMBASE. The exact details of the search are provided in Appendix A. An article was considered eligible for inclusion in this systematic review if the following criteria were met:

- Patients with lung cancer
- Original real world data/observational data (no post hoc analysis of trial data)
- Intervention under study is a systemic drug treatment
- Comparison of at least two systemic treatment options (e.g. treatment A vs treatment B, treatment A vs best supportive care (BSC), differences in timing, dose or duration of treatment A). Of note: articles were also included if systemic treatments were compared in a population that underwent concurrent surgery or radiation.

The first selection of articles (latest date was set at July 1st 2015) was screened for eligibility based on title by a single reviewer (BP). Subsequently, abstracts were independently screened for eligibility by two reviewers (BP and EvD). Finally, full text articles were examined by the same two reviewers. Consensus was sought in case of differences between reviewers. No reference tracking was done. Fig. 1 provides a flow diagram giving an overview of the search criteria and the yield at the different stages of study selection.

2.2. Data extraction

From all articles selected the following characteristics were captured: number of patients studied, first author, year of publication, study design, single centre study (yes/no), statistical method, details on covariate analysis, use of interaction terms, treatments compared, results of primary analyses and conclusion.

2.3. Quality assessment

All selected articles were quality assessed according to the Good ReSeArch for Comparative Effectiveness (GRACE) checklist. The GRACE checklist is a 11-item checklist that has been developed to assess the methodological quality and informational value of a CER study in a structured manner [4]. Dreyer et al. performed a validation using a large number of raters to determine how the individual items performed when applied to expert opinions on quality [4]. The GRACE checklist items were shown to perform better than opinions from individual experts and concurrent expert opinions.

In our study, a point could be earned when the specific GRACE item was considered sufficiently fit for purpose according to the GRACE checklist definitions (Table 2). The checklist is subdivided into six items relating to data and five relating to methods. Because the GRACE checklist is a general, not (lung) oncology specific, checklist, some criteria for quality were made oncology specific and/or less prone to subjective interpretation in a consensus meeting after reviewing the first ten studies by the two independent reviewers. Subsequently, these specified criteria were applied to all other studies. For all items in the checklist a study could score no, half or one point. A final score was calculated by summing the assigned points (range 0–11).

2.4. Statistical analysis

The average GRACE score for single centre studies compared to non-single centre studies was performed using a t-test. The statistical analysis was performed using IBM SPSS statistics (version 22.0).

3. Results

3.1. Descriptive results

The first database queries yielded 299 and 358 articles for MEDLINE and EMBASE respectively. After exclusion of duplicates, 419 of 551 articles were excluded based on the title. A total of 132 abstracts were then screened by the two reviewers, resulting in a total of 48 articles eligible for assessment of the full text. The main reason for exclusion in this step was “no comparator” (single regimen/agent studies) (44% of excluded abstracts). After the assessment of the full text articles, another ten articles were excluded for similar reasons (Fig. 1). In total, the final selection
comprised 38 comparative effectiveness research articles. Table 1 provides an overview of this final selection of articles. All studies were designed as retrospective observational cohort studies. A Cox proportional hazards model was most often used for statistical analysis (29/39). Only one study used an interaction term in their multivariate analysis. Zhu et al. looked at a treatment-by-histology interaction [51]. The vast majority (36/38) of studies involved non-small-cell lung cancer (NSCLC) of which eight studies (22%) compared the effectiveness of gefitinib with erlotinib. The median number of patients per study was 202 (range 21–10064). The first study was published in 1998 [5]. Fig. 2 shows the number of studies published every year. This number increased from one in 1998 to eight in 2014.

3.2. Quality assessment

The consensus meeting after reviewing the first ten articles resulted in specification of the GRACE checklist criteria on seven out of the 11 items. This specification is presented in Table 2 and has been applied to all other articles. When mortality was studied together with other outcomes, but without distinction between primary and secondary outcomes, mortality was considered the primary outcome.

The final scores are included in Table 1 and Fig. 2. Overall, the mean final score was 7.9 with a standard deviation of 1.0. The mean score was different for single centre studies (7.7) compared to non-single centre (8.3) studies but did not reach statistical significance (p = 0.055). Half of the studies scored eight or more points. When focussing on the individual items, least points were scored for item M5 (Fig. 3). Item M5 checks whether an evaluation of potential biased assessment of exposure or outcome (impact of varying exposure and/or outcome definitions on results) was conducted. For items D1 and D4, the maximum possible score was granted in less than 60% of the studies. Item D1 checks for adequate recording of treatment exposure and item D4 checks whether primary outcomes have been validated adequately. In contrast, all studies scored maximum on item D5 that checks whether the primary outcome was measured or identified in an equivalent manner between the treatment/intervention group and the comparison groups.

The highest score was adjudicated to Luo et al. [31] scoring ten out of 11 points because data on drug exposure (GRACE item D1),
<table>
<thead>
<tr>
<th>Reference</th>
<th>Comparators and patient group (n)</th>
<th>Statistical methods</th>
<th>Covariate analysis</th>
<th>Population</th>
<th>Results</th>
<th>Conclusion</th>
<th>Single centre</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amun 2012 [15]</td>
<td>Post-operative RT with CT (n = 15) vs without CT (n = 46)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC stage III-N2</td>
<td>OS CT vs no CT, HR 0.23 (p = 0.009)</td>
<td>Aggressive consolidative therapy may improve outcome</td>
<td>Yes</td>
<td>7.5</td>
</tr>
<tr>
<td>Boffa 2015 [16]</td>
<td>Pre- (n = 333) vs postoperative (n = 351) CT and RT * Pre- (n = 1023) vs RT * postoperative (n = 298)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC stage III-N2</td>
<td>Post-operative CT vs pre-operative CT, HR 1.05 (p = 0.44)</td>
<td>No superior CT approach could be identified</td>
<td>No</td>
<td>7.5</td>
</tr>
<tr>
<td>Brunelli 2006 [17]</td>
<td>Neoadjuvant gemcitabine-cisplatin (n = 70) vs no neoadjuvant CT (n = 70)</td>
<td>CPH</td>
<td>Propensity pair matching within cohort</td>
<td>NSCLC N2 or T4</td>
<td>ICU admission, blood transfusion, length of hospital stay, no difference</td>
<td>No superior CT approach could be identified</td>
<td>Yes</td>
<td>7.5</td>
</tr>
<tr>
<td>Cai 2014 [18]</td>
<td>WBRT/SRS/surgery plus CT with (n = 91) vs without TKI (n = 141)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC with brain metastases</td>
<td>TKI + CT may be beneficial for OS and PFS</td>
<td>Yes</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Chang 2014 [19]</td>
<td>Adjuvant pacitaxel-carboplatin (PC) (n = 207) vs vinorelbine-cisplatin (NP) (n = 231)</td>
<td>KM with log rank</td>
<td>Stratified analysis</td>
<td>NSCLC stage IIIA</td>
<td>No significant difference in OS or PFS</td>
<td>Yes</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>Chung 2011 [20]</td>
<td>Low-dose docetaxel (n = 79) vs pemetrexed (n = 100)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC stage III/IV</td>
<td>PFS: 63.6 months (PC) vs 54.8 months (NP), (p = 0.068)</td>
<td>Low dose docetaxel as tolerable as pemetrexed (authors)</td>
<td>Yes</td>
<td>8.5</td>
</tr>
<tr>
<td>Cromwell 2011 [21]</td>
<td>Erlotinib (n = 133) vs docetaxel (n = 68)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC stage III/IV</td>
<td>OS: 251 days (erlotinib) vs 310 days (docetaxel), (P = NS)</td>
<td>No significant difference in OS or PFS</td>
<td>Yes</td>
<td>9.5</td>
</tr>
<tr>
<td>Dranitsaris 2013 [22]</td>
<td>Bevacizumab maintenance (n = 74) vs no maintenance treatment (n = 198)</td>
<td>CPH</td>
<td>Landmark and propensity score</td>
<td>NSCLC stage III/IV</td>
<td>OS maintenance vs no maintenance HR 0.52 (0.37–0.73) (landmark)</td>
<td>Bevacizumab maintenance contributed to an OS benefit</td>
<td>Yes</td>
<td>9.5</td>
</tr>
<tr>
<td>Earle 2001 [23,24]</td>
<td>CT (n = 2012) vs no CT (n = 4220)</td>
<td>CPH</td>
<td>Propensity score and instrumental variable</td>
<td>NSCLC stage IV &gt;65 yrs</td>
<td>DCR: 95.6% (TKI) vs 70.9% (CT), (P = 0.032)</td>
<td>CT effective in elderly patients (with comorbid conditions)</td>
<td>No</td>
<td>8.5</td>
</tr>
<tr>
<td>Fiala 2013 [25]</td>
<td>TKI (n = 23) vs CT (n = 31)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC stage III/IV, EGFR+</td>
<td>DCR: 95.6% (TKI) vs 70.9% (CT), (P = 0.032)</td>
<td>TKI treatment was associated with significant better DCR and PFS</td>
<td>Yes</td>
<td>8</td>
</tr>
<tr>
<td>Galetta 2012 [26]</td>
<td>Neoadjuvant CT (n = 26) vs no CT (n = 21)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC stage I-III</td>
<td>OS at 5 years: 62.7% (CT) vs 10.7% (no CT), (P = 0.001)</td>
<td>Neoadjuvant CT for bronchoangioplasty interventions allows good long term outcomes</td>
<td>Yes</td>
<td>7.5</td>
</tr>
<tr>
<td>Kim 2010 [13]</td>
<td>Gefitinib (n = 171) vs erlotinib (n = 171)</td>
<td>CPH</td>
<td>Multivariate, pair matched patient selection</td>
<td>NSCLC stage III/IV</td>
<td>ORR: 38% (gefitinib) vs 32.2% (erlotinib), (P = 0.27)</td>
<td>No significant differences in outcomes</td>
<td>Yes</td>
<td>9</td>
</tr>
<tr>
<td>Study (Year)</td>
<td>Design</td>
<td>Treatment</td>
<td>Control</td>
<td>Outcomes</td>
<td>Additional Details</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------</td>
<td>--------</td>
<td>-----------</td>
<td>---------</td>
<td>----------</td>
<td>-------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ko 2007 [27]</td>
<td>Weekly (n = 18) vs triweekly (n = 19) docetaxel</td>
<td>KM with log rank</td>
<td>Univariate</td>
<td>NSCLC stage IIIB/IV</td>
<td>OS: 12.6 (gefitinib) vs 12.1 months (erlotinib) (P = 0.99); PFS: 4.6 (gefitinib) vs 2.7 months (erlotinib) (P = 0.06); Weekly tolerable and comparable activity to that of triweekly regimen</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lee 2013 [28]</td>
<td>Erlotinib (n = 14) vs gefitinib (n = 11)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC with leptomeningeal carcinomatosis</td>
<td>Cytological conversion rate: 64.3% (erlotinib) vs 9.1% (gefitinib) (P = 0.012); Erlotinib showed a better control rate for leptomeningeal carcinomatosis than gefitinib</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Liao 2015 [29]</td>
<td>Platinum combined with gemcitabine (n = 482), docetaxel (n = 143), paclitaxel (n = 114) or vinorelbine (n = 155)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC stage IIIB/IV (squamous)</td>
<td>No differences in OS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lim 2014 [30]</td>
<td>Erlotinib (n = 121) vs gefitinib (n = 121)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC recurrent or stage IIIB/IV, EGFR positive</td>
<td>ORR: 76.9% (gefitinib) vs 74.4% (erlotinib) (P = 0.58); Similar effectiveness for both TKIs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luo 2011 [31]</td>
<td>Cisplatin-based chemotherapy (n = 788) vs carboplatin-based chemotherapy (n = 1014)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC stage IIIB/IV</td>
<td>OS: 324 days (cisplatin) vs 286 days (carboplatin) (P = 0.003) in stage III; Cisplatin-based CT was associated with better OS in patients with stage IIIB No significant difference in survival</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Luo 2012 [32]</td>
<td>Vinorelbine-ifosfamide-cisplatin (NIP) (n = 80) vs etoposide-cisplatin (EP)(n = 96)</td>
<td>KM with log rank</td>
<td>Univariate</td>
<td>SCLC</td>
<td>ORR: 30.0% (NIP) vs 38.5% (EP) (P = 0.24); No significant difference in survival</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Machtay 2004 [33]</td>
<td>Neoadjuvant etoposide/cisplatin + RT (n = 22) vs carboplatin/paclitaxel + RT (n = 31)</td>
<td>KM with log rank</td>
<td>Univariate</td>
<td>NSCLC stage III</td>
<td>OS at 4 years: 36% (EP + RT) vs 26% (CP + RT) (P = 0.67); No significant difference in survival</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ng 2008 [34]</td>
<td>TKI (n = 22 2nd line, n = 31 3rd line) compared to docetaxel (n = 52 2ndline, n = 22 3rd line)</td>
<td>KM with log rank</td>
<td>Univariate</td>
<td>NSCLC stage IIIB/IV</td>
<td>OS 2nd line: 288 days (TKI) vs 136 days (docetaxel) (p = 0.23); OS 3rd line: 100 days (TKI) vs 160 days (docetaxel) (p = 0.67); Use of 2nd line TKI equivalent effectiveness as docetaxel Use of 3rd line docetaxel equivalent effectiveness as TKI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nishiyama 2015 [35]</td>
<td>Erlotinib (n = 31) vs pemetrexed (n = 66) vs docetaxel (ns = 106)</td>
<td>CPH</td>
<td>Propensity score</td>
<td>NSCLC advanced, EGFR negative</td>
<td>ORR: 7.4, 6.1 (pemetrexed) vs 9.3 (docetaxel) months (P = 0.53); No difference between treatment on PFS and OS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Popat 2008 [36]</td>
<td>Erlotinib (n = 29 2nd line, n = 23 3rd line) vs docetaxel (n = 79 2nd line and n = 20 3rd line) vs gefitinib (n = 85 2nd line and n = 53 3rd line)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC stage IIIB/IV</td>
<td>OS second line: 24 weeks (erlotinib) vs 43 weeks (docetaxel) vs 25 weeks (gefitinib) (P = 0.17); No significant difference in survival</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ritzwoller 2014 [37]</td>
<td>Bevacizumab-carboplatin-paclitaxel (BCP, n = 198) vs carboplatin-paclitaxel (CP, n = 1407)</td>
<td>CPH</td>
<td>Multivariate, propensity score</td>
<td>NSCLC stage IIIB/IV non squamous</td>
<td>OS: in propensity score adjusted model BCP was associated with significant better survival</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 1 (Continued)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Comparators and patient group (n)</th>
<th>Statistical methods</th>
<th>Covariate analysis</th>
<th>Population</th>
<th>Results</th>
<th>Conclusion</th>
<th>Single centre</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shah 2013 [38]</td>
<td>Pemetrexed platinum (PP) (n = 300) vs paclitaxel carboplatin (PC) (n = 300) vs PC bevacizumab (PCB) (n = 300) Gefitinib (n = 655) vs erlotinib (n = 329)</td>
<td>CPH</td>
<td>Unknown</td>
<td>NSCLC stage IIIB/IV or progressive disease non squamous</td>
<td>PFS: 134 (PP) vs 106 (CP) vs 126 days (P &lt; 0.001)</td>
<td>PP was associated with significant better PFS</td>
<td>No</td>
<td>9</td>
</tr>
<tr>
<td>Shao 2013 [39]</td>
<td>Pacitaxel carboplatin (PC) (n = 11) with or without bevacizumab (PCB) (n = 10)</td>
<td>KM with log rank</td>
<td>Univariate</td>
<td>NSCLC stage III with ILD non squamous</td>
<td>OS: 16,1 (PC) vs 9,7 (PC) months (p = 0,77)</td>
<td>Gefitinib and erlotinib similar effectiveness</td>
<td>No</td>
<td>8,5</td>
</tr>
<tr>
<td>Shimizu 2014 [40]</td>
<td>Single-agent CT (n = 55) vs combination CT (n = 138)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC stage IIIB/IV</td>
<td>ORR: 25.4% (combination) vs 9.1% (single) (P = 0.012)</td>
<td>Potential role of prolonging PFS using combination therapy</td>
<td>Yes</td>
<td>7,5</td>
</tr>
<tr>
<td>Shin 2011 [41]</td>
<td>Gepitinib (n = 100) vs erlotinib (n = 62)</td>
<td>KM with log rank</td>
<td>Univariate</td>
<td>NSCLC squamous</td>
<td>DCR: 40.0% (gefitinib) vs 41.4% (erlotinib) (P = 0.44)</td>
<td>Gefitinib and erlotinib similar effectiveness</td>
<td>Yes</td>
<td>8,5</td>
</tr>
<tr>
<td>Song 2013 [42]</td>
<td>Single-agent CT (n = 24) vs combination CT (n = 69) vs TKI (n = 33)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC extensive stage</td>
<td>ORR: 28.6% (single) vs 2.3% (combination) vs 3.0% (TKI) months (P = 0.03)</td>
<td>Advanced NSCLC could benefit from 3rd line treatment, mono therapy is recommended</td>
<td>Yes</td>
<td>7,5</td>
</tr>
<tr>
<td>Tanaka 1998 [44]</td>
<td>Adjuvant tegafur/uracil (UFT) (n = 98) vs no UFT (n = 557)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC stage II-IIIa</td>
<td>OS at 5 years: 76.5% (UFT) vs 58.6% (no UFT) (P = 0.005)</td>
<td>Efficacy of oral UFT was proposed adjuvant NSCLC</td>
<td>Yes</td>
<td>9,5</td>
</tr>
<tr>
<td>Tang 2014 [45]</td>
<td>Comparison of different TKI orders (n = 120)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC</td>
<td>No difference among sequence of TKI treatment</td>
<td>No difference among sequence of TKI treatment</td>
<td>Yes</td>
<td>7</td>
</tr>
<tr>
<td>Wang 2014 [46]</td>
<td>Propensity matched platinum pemetrexed (PP) pairs with various platinum doublets (total pairs n = 484) Gefitinib (n = 440) vs erlotinib (n = 276)</td>
<td>CPH</td>
<td>Multivariate, propensity score</td>
<td>NSCLC advanced Stage</td>
<td>DCR: PP had significantly better DCR compared to all other regimens</td>
<td>Superior clinical effectiveness of PP compared to other platinum based doublets</td>
<td>No</td>
<td>8</td>
</tr>
<tr>
<td>Wu 2011 [47]</td>
<td>Single-agent CT (n = 58) vs combination CT (n = 89) vs TKI (n = 61) vs CT + targeted therapy (n = 25)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC stage IIIB/IV</td>
<td>ORR: 12.8% (gefitinib) vs 13.9% (erlotinib) (P = 0.84)</td>
<td>Type of TKI not associated with treatment outcomes independent of EGFR status</td>
<td>Yes</td>
<td>8,5</td>
</tr>
<tr>
<td>Ying Geng 2013 [48]</td>
<td>Single-agent CT (n = 58) vs combination CT (n = 89) vs TKI (n = 61) vs CT + targeted therapy (n = 25)</td>
<td>CPH</td>
<td>Multivariate</td>
<td>NSCLC stage IIIB/IV</td>
<td>ORR: 12.4 (gefitinib) vs 6.8 months (erlotinib) (P = 0.16) (wild type EGFR)</td>
<td>TKI and CT plus targeted therapy showed increased OS compared with single and doublet CT</td>
<td>Yes</td>
<td>6,5</td>
</tr>
</tbody>
</table>
primary outcome (D2), and relevant covariates (D6) was ade-
quately recorded. Also, they used mortality as the primary outcome for both treatment groups (D5) which is an objective outcome (D3) that was validated by reviewing admission docu-
ments or phone calls to family members of subjects (D4). Methodologically, the study was restricted to new initiators (M1), comparing two treatment options during the same time frame (M2) using a Cox regression model (M3) to adjust for covariates age, gender, clinical stage, and histology (M3). Performance status was used as an inclusion criteria and therefore not included in the regression model. Finally, there was no risk of immortal time bias (M4) but no meaningful analysis were conducted to test key assumptions on which the primary results are based (sensitivity analysis) (M5).

4. Discussion

In this systematic review we present all CER studies that are published to date in the field of lung oncology. We observed a clear increase in the number of published CER studies in the recent years but with no increase in overall quality score. Based on this quality assessment, we identified general shortcomings that should be addressed in future CER studies to fulfill its promise to assist clinical decision-making.

In the past years, a number of overview and opinion leaders’ articles have been published on CER in oncology. These publica-
tions have addressed basic information about CER, the oppor-
tunities, methodological challenges, and future perspectives [6–9]. Although recommendations about study design, methodology and data collection are often made [8,9], our study is the first to systematically review the currently available CER studies regarding these quality criteria in the field of lung oncology. Our review showed that a considerable part of the published CER studies so far did not score maximum on several quality criteria.

To start, we observed that in a high percentage (26/38) of the published CER studies, patient selection and data collection was performed retrospectively through medical chart review. Such strategy holds high risk for introducing both selection and information bias, something that cannot be adjusted for later on. Second, sensitivity analyses to explore the robustness of the study findings were absent in almost all studies selected (32/38).

Finally, and very importantly, the set of potential confounders explored in the analyses was rather limited in more than half of the studies. As main example, patients’ performance status was not accounted for in 12 out of 38 studies as potential confounding factor, while it is obvious that performance status at start of treatment could have a strong relation with both selection of therapy as well as clinical outcome. The reason for not including performance status was in most cases unavailability of that information.

There are some limitations to our study that need to be addressed. First, although the GRACE checklist was developed for quality assessment of observational CER studies [4], it is not oncology specific. This made us tailor the checklist towards oncology specific study characteristics for unambiguous interpre-
tation (Table 2). In addition, the GRACE score does not identify studies fit for purpose and those not. It mainly provides a single quantitative summary score for data quality and methodology. Finally, the overall GRACE score appeared not to discriminate very well between quality of studies (all scores were in the range of 6–10 with small standard deviation of 1.0), but did, however, succeed to identify individual items that scored negative often. In accordance with our observation, Dreyer et al. also demonstrated the GRACE checklist did not to achieve clear discrimination between studies fit for purpose and those not in their testing and validation effort [4]. Nevertheless, the GRACE checklist is the only available tool designed to assess study quality of specifically observational CER. Other guidelines such as The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) recommendations provide guidance on the reporting of observa-
tional studies to facilitate critical appraisal and interpretation of results, rather than assessing study quality [10]. Moreover, in contrast to other guidelines/tools, the GRACE score has been tested for validity [4].

Despite the identified common methodological shortcomings in studies captured in this review, we think there are some promising developments in (lung) oncology that could help to overcome these shortcomings in the near future. A major development is the construction of so-called learning healthcare systems in clinical oncology [11,12]. One important aspect of such systems is that relevant data collection is seamlessly embedded in the delivery of care. This this will boost the quality of the data and
## Table 2
GRACE checklist items partially tailored towards lung oncology.

<table>
<thead>
<tr>
<th>Item</th>
<th>GRACE description</th>
<th>Rationale</th>
<th>Criterion</th>
<th>Points</th>
</tr>
</thead>
<tbody>
<tr>
<td>D1</td>
<td>Treatment and/or important details of treatment exposure adequately recorded for the study purpose in the data source(s)?</td>
<td>When comparing different treatments, data on drug doses are essential and should therefore be available.</td>
<td>Description of doses is missing.</td>
<td>0</td>
</tr>
<tr>
<td>D2</td>
<td>Were the primary outcomes adequately recorded for the study purpose (e.g., available in sufficient detail through data source(s))?</td>
<td>Medical hospital records are considered less accurate than prospectively collected data or insurance records [53]</td>
<td>Database limited to medical records</td>
<td>0.5</td>
</tr>
<tr>
<td>D3</td>
<td>Was the primary clinical outcome(s) measured objectively rather than subject to clinical judgment (e.g., opinion about whether the patient’s condition has improved)?</td>
<td>Mortality is the most objective outcome and progression is less objective.</td>
<td>Progression measured by RECIST criteria</td>
<td>0.5</td>
</tr>
<tr>
<td>D4</td>
<td>Were primary outcomes validated, adjudicated, or otherwise known to be valid in a similar population?</td>
<td>The validation of the primary outcome depends on database that was used.</td>
<td>Progression measured by a non-validated system</td>
<td>0</td>
</tr>
<tr>
<td>D5</td>
<td>Was the primary outcome measured or identified in an equivalent manner between the treatment/intervention group and the comparison groups?</td>
<td>Yes</td>
<td>No, or not enough information in article</td>
<td>0</td>
</tr>
<tr>
<td>D6</td>
<td>Were important covariates that may be known confounders or effect modifiers available and recorded?</td>
<td>Performance status is the most important covariate in the oncology [54]</td>
<td>Performance status is not recorded and there is no restriction of certain performance status values in the inclusion criteria</td>
<td>0</td>
</tr>
<tr>
<td>M1</td>
<td>Was the study (or analysis) population restricted to new initiators of treatment or those starting a new course of treatment?</td>
<td>Yes, only new initiators of the treatment of interest were included in the cohort, or for surgical procedures and devices, including only patients who never had the treatment before the start of study follow-up.</td>
<td>No, or not enough information in article</td>
<td>0</td>
</tr>
<tr>
<td>M2</td>
<td>If 1 or more comparison groups were used, were they concurrent comparators? If not, did the authors justify the use of historical comparison groups?</td>
<td>Yes, data were collected during the same time period as the treatment group (“concurrent”), or historical comparators used with reasonable justification</td>
<td>No, historical comparators used without being scientifically justifiable, or not enough information in article.</td>
<td>0</td>
</tr>
<tr>
<td>M3</td>
<td>Were important covariates, confounding and effect modifying variables taken into account in the design and/or analysis?</td>
<td>Using appropriate statistical methods to adjust for important covariates is considered essential for the quality of a CER study.</td>
<td>Performance status or another important covariate is not measured but a multivariate analysis is performed. Only stratification or restriction is performed.</td>
<td>0.5</td>
</tr>
<tr>
<td>M4</td>
<td>Is the classification of exposed and unexposed person-time free of “immortal time bias”?</td>
<td>Yes</td>
<td>No, or not enough information in article</td>
<td>1</td>
</tr>
<tr>
<td>M5</td>
<td>Were any meaningful analyses conducted to test key assumptions on which primary results are based?</td>
<td>Sensitivity analysis is not mentioned.</td>
<td>No, or not enough information in article</td>
<td>0</td>
</tr>
</tbody>
</table>

* Tailored towards lung oncology.
prevent information bias through missing data for CER studies. A standard set of relevant patient characteristics will become available independent of the interventions under study further preventing selection bias. Possibly, future CER studies based on data from these systems will score positive on most if not all of the quality criteria discussed in this review. In the long run, this can really pave the way to incorporate CER studies in clinical decision-making and evaluate its merits on patient outcome. To investigate the influence of a learning healthcare system on the quality of CER studies, we compared the average GRACE score of single centre studies to non-single centre studies. The latter group consisted of studies performed in national cancer databases/registries or collaborating medical centres. These platforms are generally established to embrace the concept of a learning healthcare system and are more likely to perform higher quality observational CER studies. This is reflected in the average GRACE score of non-single centre studies that was higher compared to single centre studies, although statistical significance was not reached ($p=0.055$).

When looking at current treatment guidelines for NSCLC, we were able to identify only one CER study from this systematic review being incorporated [13,14]. Why do guidelines in clinical oncology make little or no use of observational CER of systemic therapies? Barriers for using observational CER in treatment guidelines may include (I) quality concerns related to an observational study design, (II) unavailability of (oncology specific) tools for quality assessment, and (III) unavailability of a composite evidence grading system for recommendations based on data from both observational studies and RCTs. We think that the utilization of concurrent evidence has the potential to strengthen treatment recommendations because observational CER is able to answer questions that relate to the external validity of results from RCTs.
Moreover, the rapid development of new medical treatments will give rise to many more clinical questions that need to be answered but are financially and practically impossible to answer with RCTs only.

Despite the growing interest in CER studies in the recent years, it can be concluded that most of the CER studies in lung oncology published up to now share methodological shortcomings possibly limiting their applicability in clinical decision-making. However, many recent developments in lung oncology practice could help to overcome these shortcomings in the near future. Until then, RCTs will remain provisionally the standard for comparing treatments.

Conflicts of interest

None.

Acknowledgement

The authors thank Carla Sloof, clinical librarian, for literature search assistance.

Appendix A.

MEDLINE


References