Adv Ther (2021) 38:1843–1859
https://doi.org/10.1007/s12325-021-01659-0

ORIGINAL RESEARCH

Characterization of a Real-World Response Variable and Comparison with RECIST-Based Response Rates from Clinical Trials in Advanced NSCLC

Xinran Ma · Lawrence Bellomo · Kelly Magee · Caroline S. Bennette · Olga Tymejczyk · Meghna Samant · Melisa Tucker · Nathan Nussbaum · Bryan E. Bowser · Joshua S. Kraut · Ariel Bulua Bourla

Received: December 2, 2020 / Accepted: February 8, 2021 / Published online: March 5, 2021

ABSTRACT

Introduction: Effectiveness metrics for real-word research, analogous to clinical trial ones, are needed. This study aimed to develop a real-world response (rwR) variable applicable to solid tumors and to evaluate its clinical relevance and meaningfulness.

Methods: This retrospective study used patient cohorts with advanced non-small cell lung cancer from a nationwide, de-identified electronic health record (EHR)-derived database. Disease burden information abstracted manually was classified into response categories anchored to discrete therapy lines (per patient-line). In part 1, we quantified the feasibility and reliability of data capture, and estimated the association between rwR status and real-world progression-free survival (rwPFS) and real-world overall survival (rwOS). In part 2, we investigated the correlation between published clinical trial overall response rates (ORRs) and real-world response rates (rwRRs) from corresponding real-world patient cohorts.

Results: In part 1, 85.4% of patients (N = 3248) had at least one radiographic assessment documented. Median abstraction time per patient-line was 15.0 min (IQR 7.8–28.1). Inter-abstractor agreement on presence/absence of at least one assessment was 0.94 (95% CI 0.92–0.96; n = 503 patient-lines abstracted in duplicate); inter-abstractor agreement on best confirmed response category was 0.82 (95% CI 0.78–0.86; n = 384 with at least one captured assessment). Confirmed responders at a 3-month landmark showed significantly lower risk of death and progression in rwOS and rwPFS analyses across all line settings. In part 2, rwRRs (from 12 rw cohorts) showed a high correlation with trial ORRs (Spearman's ρ = 0.99).

Conclusions: We developed a rwR variable generated from clinician assessments documented in EHRs following radiographic evaluations. This variable provides clinically meaningful information and may provide a real-world measure of treatment effectiveness.

Keywords: Real-world data; Real-world evidence; RECIST; Response; RWD; RWE

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12325-021-01659-0.

X. Ma (✉) · L. Bellomo · K. Magee · C. S. Bennette · O. Tymejczyk · M. Samant · M. Tucker · N. Nussbaum · B. E. Bowser · J. S. Kraut · A. B. Bourla
Flatiron Health, Inc, New York, NY, USA
e-mail: xma@flatiron.com

N. Nussbaum
New York University School of Medicine, New York, NY, USA

Key Summary Points

Determining the occurrence of "tumor responses" in cohorts of real-world patients in oncology research requires the development of variables suited to be applied to real-world data sources, such as electronic health records (EHRs).

We describe the development of a real-world response variable that can be derived from clinicians' assessments documented in the EHR after radiographic evaluations in patients with advanced non-small cell lung cancer.

This variable can be extracted in a feasible and reliable fashion, and provides a measure of treatment effectiveness, as shown by correlations of the associated endpoint (real-world response rate) with other clinically meaningful endpoints (such as real-world progression-free and overall survival), as well as with clinical trial results obtained in matching clinical settings.

Future research will be needed to investigate potential expansions of the use of this variable to other solid tumor settings, and to better understand the relationship with tumor response as determined by clinical trial criteria.

DIGITAL FEATURES

This article is published with digital features, including a summary slide, to facilitate understanding of the article. To view digital features for this article, go to https://doi.org/10.6084/m9.figshare.13721353

INTRODUCTION

The use of health data collected during routine care (real-world data, RWD) has broadened in recent years, expanding the possibilities for observational studies and outcomes research [1]. Analyzing RWD to generate high-quality real-world evidence (RWE) has emerged as a potential complement to traditional clinical trials; incorporating this type of research to inform drug development, or contribute to regulatory decision-making, represents a compelling prospect.

With the advent of digitization across healthcare delivery systems [2], electronic health records (EHRs) are becoming a key RWD source. Developing appropriate quality and analytic benchmarks for EHR-derived RWD is a critical step in the generation of interpretable RWE fit to support decision-making during clinical development or regulatory processes, while upholding the standards required to preserve patient well-being [3].

In oncology clinical trials, improvements in overall survival are a key measure of an intervention's efficacy, and one of the gold standards used in regulatory approvals [4]. However, therapeutic effects either known or reasonably likely to predict clinical benefit, and measurable earlier than survival improvements, such as increases in progression-free survival (PFS) time or in durable response rates (RR), have also supported approvals [5, 6]. In traditional trials, the use of standardized criteria, such as the Response Evaluation Criteria in Solid Tumors (RECIST), helps ensure the objectivity, validity, and reliability of these outcome measurements and associated endpoints [7]. However, the application of RECIST requires imaging evaluations that are thorough, standardized, and longitudinally consistent, while routine care does not necessarily follow clinical trial standards and radiographic images are rarely available in EHR data; these realities pose a challenge for the generation of RWE [8]. Prior work from a related group showed that retrospective application of RECIST to a large EHR-derived dataset of patients with advanced non-small cell lung cancer (aNSCLC) was not feasible, largely because of the lack of all required documentation [8]. As an alternative, that author team derived a real-world progression (rwP) variable by abstracting clinicians' notes contained in EHRs; on the basis of this variable, the authors produced scalable, well-

characterized, and clinically meaningful real-world PFS (rwPFS) results [9].

In this study, we aimed to develop a method to capture the clinician's interpretation of radiographic assessments of solid tumor burden from information contained in EHRs in order to derive a real-world response (rwR) variable, and to assess the validity of this variable and the associated endpoint (real-world response rate; rwRR).

METHODS

Overall Design and Objectives

This retrospective exploratory study was conducted in two parts: part 1 focused on the extraction of a response variable from EHR information, and on the validity assessment of that variable based on feasibility, reliability, and correlation with downstream clinical outcomes; part 2 focused on the descriptive comparison of endpoint results between ORR reported in clinical trials and rwRR obtained using this response variable in cohorts corresponding to the specific trial criteria.

Data Source

The Flatiron Health database is a nationwide longitudinal, de-identified EHR-derived database comprised of de-identified patient-level structured and unstructured data, curated via technology-enabled abstraction [10]. During the study period, this database included de-identified data from approximately 280 US cancer clinics (ca. 800 sites of care).

Institutional review board (IRB) approval of the study protocol, with a waiver of informed consent, was obtained prior to study conduct, and covers the data from all sites represented. Approval was granted by the WCG IRB. ("The Flatiron Health Real-World Evidence Parent Protocol", Tracking # FLI1-18-044).

Cohort Selection

Part 1 included a cohort of 3248 patients sourced from a database where patients had undergone next generation sequencing (NGS) testing of their tumor samples, although molecular testing was not a component of our evaluations. Patients were included if they had a diagnosis of aNSCLC (stage IIIB/IIIC/IV at initial diagnosis or recurrent advanced disease) at age at least 18 years before July 1, 2018, at least two EHR-documented clinical visits on or after January 1, 2011, and received at least one line of therapy. Patients who had incomplete historical treatment data (i.e., no structured activity within 90 days after the advanced diagnosis date) or who died within 30 days of starting first-line therapy were excluded (Supplementary Fig. 1).

In part 2, patients were drawn from a cohort of patients without NGS testing requirements, applying largely similar general selection criteria as in part 1. In addition, real-world cohorts in part 2 were aligned for consistency with 12 corresponding trial control or experimental arms, for a total of 1224 patients, of which 962 had at least one assessment documented in the EHR evaluating changes in their burden of disease. Application of trial eligibility criteria depended on data availability in the EHR (detailed in Supplementary materials).

Development of the Real-World Response (rwR) Variable

We defined response as a clinician's assessment of change in disease burden in an individual patient during a line of therapy (i.e., per patient-line). We considered only the assessments of response that followed radiographic imaging tests performed to evaluate disease burden, and the dates of those assessments were marked as "assessment time points." This information was abstracted from the EHR by trained professionals [10].

Our approach followed two steps (Fig. 1):

1. EHR document review. EHR data were curated using abstraction procedures to retrospectively capture assessment time points indexed to a line of treatment (Supplementary materials). Imaging tests performed within 2 weeks were considered one single assessment time point, dated with

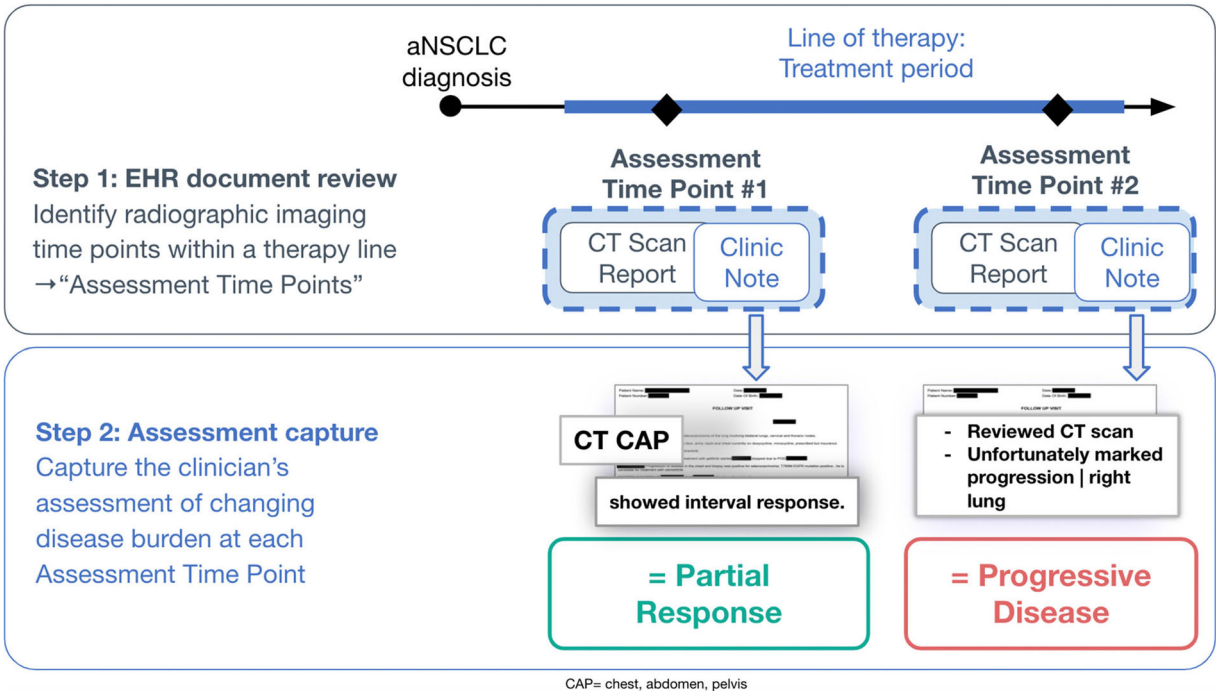

Fig. 1 Schematic representation of the steps in the rwR derivation process

the earliest imaging test date at each point. Assessment time points within 30 days following therapy initiation were not captured to allow for sufficient treatment exposure for potential impact on tumor burden dynamics.

2. Response assessment capture. At each assessment time point, abstractors captured whether any disease burden changes were reflected in the clinician's documentation, and classified the type of change using predefined categories (Table 1). Multiple

Table 1 Response categories for the real-world variable

Response category	Clinician's notes content
Real-world complete response (rwCR)	Complete resolution of disease
Real-world partial response (rwPR)	Partial reduction in tumor burden in some or all areas without any areas of increasing disease. rwPR captures a decrease in disease burden though disease is still present
Real-world stable disease (rwSD)	No change in overall disease burden. rwSD is also used to capture mixed response (some lesions increased, some lesions decreased)
Real-world progressive disease (rwPD)	Increase in disease and/or presence of any new lesions
Real-world pseudoprogression	Mentions of pseudoprogression or related terminology (e.g., tumor flare) with regard to the response scan in the setting of an immunotherapy
Indeterminate response	Explicitly stating that they are not able to make a response assessment determination
Not documented	Record of a response scan but no record of a clinician assessment of the scan

assessment time points were abstracted when present.

Statistical Analyses

Feasibility

The feasibility of abstracting rwR from a large EHR-derived database was determined by the completeness of data capture and the abstraction time length. To assess the completeness of data capture, we summarized the percentage of patients with a radiographic assessment documented in the EHR on at least one line of therapy, and the percentage of patients who had a documented clinician's interpretation of change in disease burden within the same line.

Time to first assessment time point, and observed frequency of assessments were summarized. We estimated the probability of patients having a radiographic assessment within 3 or 6 months after initiating a therapy line (first, second, or third line of therapy) using the Kaplan–Meier method. Patients were censored at the time of death, loss to follow-up, or 1 day before the initiation of subsequent therapy, whichever occurred earliest. We reported the median observed frequency of assessment for each line setting among patients with two or more assessments performed up to and including the last assessment or first documented progressive disease (PD) event, whichever occurred first. Observed frequency of assessment was calculated on a patient-by-patient basis as the average time between consecutive assessments during the line of interest.

Abstraction time per patient chart was computed by the abstractor software interface and assessed descriptively.

Reliability

Reliability was assessed by estimating inter-abstractor agreement rates among a random sample of 503 duplicately abstracted patient-lines (468 distinct patients). For each patient-line, we reported the proportion where abstractors agreed on the presence and absence of at least one radiographic assessment, and the percentage agreement on the best response category among those with presence of at least one clinician assessment. In addition to patient-line level agreement rates, the proportion of assessment time points where abstractors agreed on the dates and the response category was reported.

Analysis of Real-World Endpoints

In order to assess the clinical relevance of the rwR variable, we analyzed several endpoints in the study cohort. rwRR was defined as the proportion of patients with a complete or partial response determination (i.e., responders) among patients with at least one known assessment on a given treatment line. Since routine care tends to lack dedicated radiographic assessments to confirm a response (as required in RECIST), we considered "confirmed responses" those with an assessment in direct succession after the initial response indicating "stable disease" or better, we interpreted that as documentation of a continued response. rwRR with confirmation (rwRRconf) was defined as the proportion of confirmed responders on a given treatment line.

Real-world overall survival (rwOS, based on a composite mortality variable generated by aggregation of RWD sources) and real-world progression-free survival (rwPFS) were derived as previously described [8, 11]. The index date was the start of therapy of interest.

Clinical Relevance

Clinical relevance of rwR was assessed in two phases:

1. *Association with rwPFS and rwOS*—In part 1, the association between response and other clinically relevant events (progression and mortality) was evaluated using Cox proportional hazard modeling in landmark analyses (at 3 and 6 months) to compare rwOS and rwPFS between responders and non-responders (Supplementary materials). Patients who had at least one response assessment documented in the EHR and were at risk of the relevant events were included in the analysis. Multivariable analyses were adjusted for age at advanced diagnosis, smoking status, histology, and stage at initial diagnosis. We further

examined the associations in a sensitivity analysis stratified by line setting and therapy class (Supplementary materials).

2. *Comparison of rwRR to published trial results*—In part 2, rwRR results were benchmarked to cohort-level published clinical trial results obtained in comparable treatment and disease settings (Supplementary materials). The final analysis sample in this comparison consisted of 7 trials for a total of 12 treatment cohorts (7 experimental, 5 control arms) [12–18].

For each trial, we applied the eligibility criteria from the corresponding protocols to build real-world patient cohorts (based on available EHR-derived data), in addition requiring one or more documented response assessment and initiation of treatment of interest at least 6 months before the part 2 cohort data cutoff date (October 1, 2018). To optimize the relevance of the real-world cohorts, inverse odds weights were applied using published summary baseline characteristics of the trial populations [19–22]. Confidence intervals (CI) of rwRR and rwRRconf were computed using the Clopper–Pearson method. We calculated absolute differences in response rates between clinical trials and real-world cohorts and estimated the strength of the association across all comparisons using a correlation coefficient (Spearman's rho). All analyses were performed in R 3.6.1.

RESULTS

Part 1 of the study included 3248 patients (Table 2, Supplementary Fig. 1), with a median follow-up after advanced diagnosis of 14.7 months (interquartile range [IQR] 8.2–26.9).

Feasibility

Within the part 1 cohort, there were 2775 patients (85.4%) with a radiographic assessment documented in the EHR on at least one line of therapy; 2727 had at least one assessment of change in disease burden documented by the clinician within the same lines and the remaining 48 patients had not. Demographic and clinical characteristics were largely similar between patients with and without an EHR-documented radiographic assessment, except the year of advanced diagnosis. Patients with more recent advanced diagnoses were more likely to lack available assessments possibly because of shorter follow-up. The probability that a patient's first set of imaging assessment(s) occurred within a 3-month window of the initiation of first, second, and third-line therapy was 78%, 78%, and 72%, respectively, and 95%, 95%, 94% for a 6-month window.

Across patients with multiple radiographic assessments during a line of therapy, the median of the observed frequency of assessment ranged from 2 to 3 months in different line settings (Fig. 2).

The median time spent by abstractors to extract response information from the EHR was 15.0 min (IQR 7.8–28.1) per patient-line.

Reliability

In the duplicate abstraction exercise (n = 468 patients, 503 patient-lines), we estimated agreement rate for patient-line-level and time-point-level metrics. At the patient-line level (n = 474 patient-lines), in 94% (95% CI 92–96%) of cases the two abstractors agreed on the presence or absence of at least one radiographic assessment. For those with agreement on the presence of at least one assessment (n = 384 patient-lines), agreement for best response category was 81% (95% CI 76–85%) without confirmation, and 82% (95% CI 78–86%) with confirmation (Supplementary Table 2, for overall distribution of best response, Supplementary Fig. 2). At the time-point level (n = 2224), when both abstractors agreed on the presence of a response assessment on the same date (86% [95% CI 85–88%]), they agreed on the response category in 77% (95% CI 74–79%) of cases. The result was similar (77% [95% CI 74–79%]) when both abstractors agreed on the presence of a response assessment within 30 days (91% [95% CI 89–92%]).

Table 2 Baseline characteristics in the cohort for the part 1 of the study

	Part 1 cohort, *N* = 3248
Age at advanced diagnosis, median [IQR]	67.0 [60.0; 74.0]
Age at advanced diagnosis	
19–34	16 (0.5)
35–49	163 (5.0)
50–64	1147 (35.3)
65–74	1187 (36.5)
75+	735 (22.6)
Year of advanced diagnosis, *n* (%)	
2014 or prior	780 (24.0)
2015–2017	2213 (68.1)
2018	255 (7.9)
Sex, *n* (%)	
Female	1611 (49.6)
Male	1637 (50.4)
Histology, *n* (%)	
Non-squamous cell carcinoma	2485 (76.5)
NSCLC histology NOS	134 (4.1)
Squamous cell carcinoma	629 (19.4)
History of smoking, *n* (%)	
Yes	2660 (81.9)
No	575 (17.7)
Unknown/not documented	13 (0.4)
Stage at diagnosis, *n* (%)	
I	270 (8.3)
II	199 (6.1)
III	640 (19.7)
IV	2089 (64.3)
Not reported or occult	50 (1.5)
Region, *n* (%)	
North Central	452 (13.9)
Northeast	547 (16.8)
South	1683 (51.8)
Unknown	95 (2.9)

Table 2 continued

	Part 1 cohort, *N* = 3248
West	471 (14.5)
Practice type, *n* (%)	
Academic	85 (2.6)
Community	3163 (97.4)
Race, *n* (%)	
Asian	98 (3.0)
Black or African American	218 (6.7)
Other race	307 (9.5)
Unknown	266 (8.2)
White	2359 (72.6)
Vital status, *n* (%)	
Alive	1195 (36.8)
Dead	2053 (63.2)
First-line therapy class, *n* (%)	
ALK inhibitors	73 (2.2)
Anti-VEGF-based therapies	607 (18.7)
Clinical study drug-based therapies	74 (2.3)
EGFR TKIs	349 (10.7)
EGFR antibody-based therapies	14 (0.4)
Non-platinum-based chemotherapy combinations	4 (0.1)
Other therapies	15 (0.5)
PD-1/PD-L1-based therapies	568 (17.5)
Platinum-based chemotherapy combinations	1402 (43.2)
Single agent chemotherapies	142 (4.4)
PD-L1 status, *n* (%)	
No interpretation given in report	581 (17.9)
Not tested	1711 (52.7)
PD-L1 equivocal	2 (0.1)
PD-L1 negative/not detected	619 (19.1)
PD-L1 positive	188 (5.8)
Results pending	36 (1.1)
Unknown	39 (1.2)

Table 2 continued

	Part 1 cohort, *N* = 3248
Unsuccessful/indeterminate test	72 (2.2)
EGFR mutation status, *n* (%)	
Mutation negative	2720 (83.7)
Mutation positive	415 (12.8)
Not tested	94 (2.9)
Results pending	2 (0.1)
Unknown	5 (0.2)
Unsuccessful/indeterminate test	12 (0.4)
Number of lines of therapy, *n* (%)	
Received only 1 line of therapy	1150 (35.4)
Received only 2 lines of therapy	1031 (31.7)
Received only 3 lines of therapy	587 (18.1)
Received more than 3 lines of therapy	480 (14.8)

ALK anaplastic lymphoma kinase, *EGFR* epidermal growth factor receptor, *NOS* not otherwise specified, *NSCLC* non-small cell lung cancer, *PD-(L)1* programmed cell death (ligand) 1, *TKI* tyrosine kinase inhibitor, *VEGF* vascular endothelial growth factor

Association with Progression and Mortality

We used 3- and 6-month landmarks to compare rwOS and rwPFS in responders vs non-responders, with and without the requirement for rw confirmation (Table 3). For rwOS, response at 3 months was significantly associated with a decreased death risk compared to non-response across all line settings; consistent results were found when response confirmation was required. For rwPFS, responders at 3 months also had a significantly lower risk of progression or death compared to non-responders in the second-line setting (HR 0.69, 95% CI 0.59–0.81), but not in the first-line or third-line settings (HR 1.00, 95% CI 0.91–1.10; HR 0.89, 95% CI 0.67–1.18, respectively). However, when requiring confirmation, responders at 3 months had a lower risk of progression than non-responders in all line settings (HR 0.70, 95% CI 0.63–0.77; HR 0.52, 95% CI 0.43–0.63; HR 0.62, 95% CI 0.43–0.88; for first-line, second-line, and third-line respectively). Results were similar using 6 months as landmark time. The associations between response status and rwOS and rwPFS remained in multivariable Cox models, adjusting for age, stage at initial diagnosis, histology, and smoking status (Table 3). Similar associations were also found when we stratified by line and therapy class, although confidence intervals were wide for some therapy classes (Supplementary Fig. 3).

Benchmarking of rwR-Based Endpoints to RECIST-Based Trial ORRs

We generated rwRRs (unweighted, weighted, and confirmed, as described in "Methods") for 12 real-world patient cohorts aligned with registrational clinical trial cohorts (Table 4, Supplementary Fig. 1). Weighted and unweighted rates were largely similar. Comparison of rwRR or rwRRconf ("confirmed" rwR was used as per

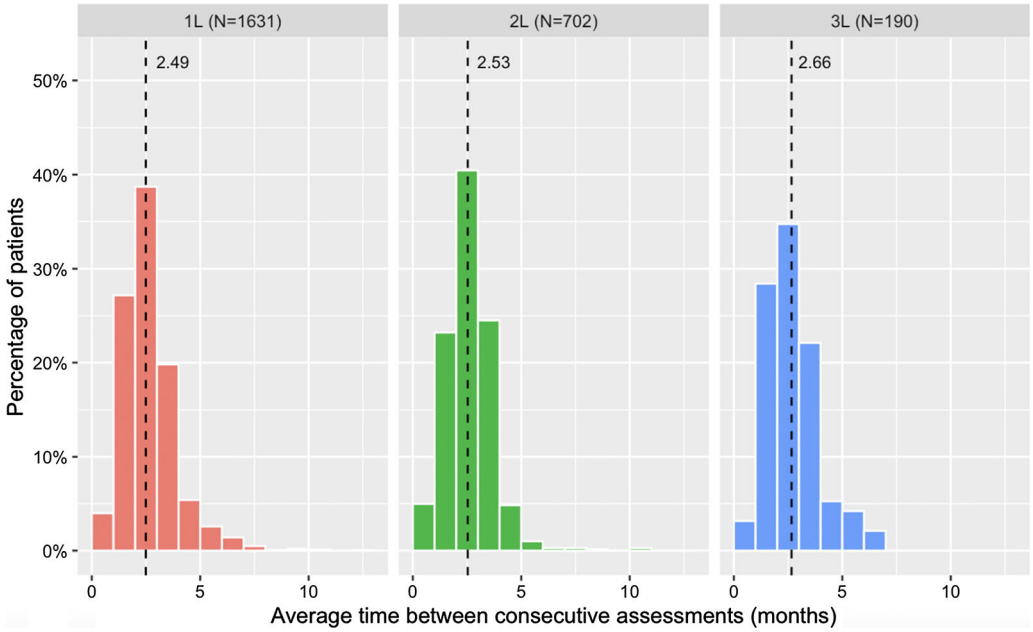

Fig. 2 Observed frequency of assessment by line of therapy for patients who had multiple assessments. Included all assessments performed up through the last assessment or the first tumor assessment that reports progressive disease, whichever came first

the corresponding trial protocol specifications, all except ALEX and AURA-3) to ORR is shown in Fig. 3. No systematic pattern appeared where rwRRs were consistently lower or higher than clinical trial ORRs, rwRRs were greater in 3 of 12 cohorts, and lower in 9; the median difference (IQR) across all 12 cohorts was − 1.7% (− 4.0 to − 0.6%). The comparison based on the investigational arm for KEYNOTE-021 registered the greatest difference: 43.7% (95% CI 34.3–53.5%) for the real-world cohort vs 55.0% (95% CI 41.6–67.9%) for the trial. Overall, the Spearman's ρ of 0.99 indicated a strong monotonic relationship between rwRRs and trial ORRs.

DISCUSSION

This study used a large real-world database of patients with aNSCLC to develop a rwR variable derived by abstracting information from EHRs, namely the clinician assessments informed by radiographic tests. We found our method to be feasible with completeness (more than 90% of patients had response assessments within 6 months of initiating therapy) and frequency in the captured assessments (2–3 month intervals) largely consistent with guideline recommendations [23]. The reliability of our approach was in line with acceptable standards, since the rates of inter-abstractor agreement, approximately 82% for best responses, were comparable to inter-observer agreement rates reported with RECIST [24–26]. Finally, the study showed that rwR is associated with rwPFS and rwOS, and demonstrated high correlation of rwRR and cohort-level ORR from clinical trials in comparable settings.

The development and characterization of endpoints for real-world clinical oncology research is the focus of ongoing efforts [8, 27–31]. We undertook an approach similar to a previously reported derivation of a progression variable from EHRs [8], with the difference of anchoring the event identification only to the documentation of radiographic evaluations. By abstracting response from the clinician's notes and assessments, our approach likely provides a more holistic synthesis of patients' clinical status than approaches leveraging radiology images or radiology reports alone [7, 29, 31].

Table 3 Cox regression on rwOS and rwPFS for responders vs non-responders with and without rw confirmation

Endpoint	Landmark	Therapy line	*n*	rwR confirmation required	
				No, HR (95% CI)	Yes, HR (95% CI)
Univariable analysis					
rwOS	3-month	Line 1	2438	0.78 (0.71, 0.87)	0.68 (0.60, 0.76)
		Line 2	1228	0.57 (0.48, 0.68)	0.47 (0.38, 0.58)
		Line 3	393	0.74 (0.54, 1.00)	0.56 (0.37, 0.86)
	6-month	Line 1	2168	0.61 (0.55, 0.68)	0.60 (0.54, 0.68)
		Line 2	1070	0.49 (0.41, 0.59)	0.46 (0.37, 0.56)
		Line 3	322	0.47 (0.34, 0.66)	0.44 (0.30, 0.65)
rwPFS	3-month	Line 1	1934	1.00 (0.91, 1.10)	0.70 (0.63, 0.77)
		Line 2	818	0.69 (0.59, 0.81)	0.52 (0.43, 0.63)
		Line 3	256	0.89 (0.67, 1.18)	0.62 (0.43, 0.88)
	6-month	Line 1	1325	1.00 (0.87, 1.14)	0.85 (0.76, 0.96)
		Line 2	537	0.80 (0.66, 0.97)	0.67 (0.55, 0.82)
		Line 3	148	0.96 (0.67, 1.40)	0.75 (0.52, 1.09)
Multivariable[a] analysis					
rwOS	3-month	Line 1	2438	0.73 (0.66, 0.82)	0.65 (0.58, 0.73)
		Line 2	1228	0.56 (0.47, 0.67)	0.47 (0.38, 0.58)
		Line 3	393	0.77 (0.56, 1.05)	0.59 (0.38, 0.90)
	6-month	Line 1	2168	0.59 (0.53, 0.66)	0.59 (0.52, 0.67)
		Line 2	1070	0.48 (0.40, 0.58)	0.46 (0.37, 0.57)
		Line 3	322	0.48 (0.34, 0.67)	0.46 (0.31, 0.67)
rwPFS	3-month	Line 1	1934	0.92 (0.83, 1.01)	0.65 (0.59, 0.72)
		Line 2	818	0.69 (0.58, 0.80)	0.52 (0.43, 0.62)
		Line 3	256	0.85 (0.64, 1.15)	0.60 (0.42, 0.86)
	6-month	Line 1	1325	0.94 (0.82, 1.07)	0.82 (0.73, 0.92)
		Line 2	537	0.80 (0.66, 0.97)	0.67 (0.55, 0.82)
		Line 3	148	0.95 (0.65, 1.41)	0.77 (0.52, 1.12)

CI confidence interval, *HR* hazard ratio, *rwOS* real-world overall survival, *rwPFS* real-world progression-free survival, *rwR* real-world response

[a] Adjusted for age at advanced diagnosis, smoking status, histology, and stage at initial diagnosis

Using this real-world variable, achieving responder/non-responder status appeared to be associated with downstream rwOS to an extent comparable to the associations reported between RECIST-based response and OS in NSCLC clinical trials [32–36]. Our

Table 4 Part 2 rwRR analysis results in specific clinical settings corresponding with registrational trials

Benchmark trial	Clinical setting/treatment	rwRR, % (95% CI)			
		N	Unweighted[a]	Weighted[b]	Weighted, confirmed[c]
KEYNOTE-024	1L IO, PD-L1 + / pembrolizumab	72	51.4 (40.1, 62.6)	54.0 (41.9, 65.7)	42.5 (31.1, 54.8)
ALEX	1L targeted, *ALK rearrang/* alectinib	60	81.7 (70.1, 89.4)	81.2 (69.1, 89.3)	67.6 (54.5, 78.4)
	1L targeted, *ALK rearrang/* crizotinib	145	64.1 (56.1, 71.5)	65.8 (51.5, 77.7)	48.1 (34.5, 62.0)
KEYNOTE-021	1L IO + chemo/ pembrolizumab + carboplatin + pemetrexed	121	66.1 (57.3, 73.9)	68.8 (59.0, 77.1)	43.7 (34.3, 53.5)
	1L chemo/ carboplatin + pemetrexed	83	63.9 (53.1, 73.4)	63.3 (51.8, 73.5)	37.6 (27.3, 49.2)
CheckMate 057	2L + IO, non-squam/ nivolumab	83	27.7 (19.2, 38.2)	27.6 (18.4, 39.3)	17.6 (10.4, 28.4)
	2L + chemo, non-squam/ docetaxel	97	23.7 (16.4, 33.1)	25.8 (17.4, 36.5)	10.9 (5.7, 19.8)
CheckMate 017	2L + IO, squam/ nivolumab	86	44.2 (34.2, 54.7)	45.7 (35.1, 56.7)	29.5 (20.6, 40.4)
	2L + chemo, squam/ docetaxel	53	24.5 (14.9, 37.6)	18.2 (9.7, 31.5)	9.8 (4.1, 21.6)
OAK	2L + IO/ atezolizumab	58	19.0 (10.9, 30.9)	25.3 (15.1, 39.1)	11.1 (4.9, 23.0)
	2L + chemo/ docetaxel	117	22.2 (15.6, 30.6)	21.8 (15.1, 30.4)	12.3 (7.4, 19.8)
AURA-3	2L + targeted, *EGFRmt/* osimertinib	97	56.7 (46.8, 66.1)	62.2 (47.7, 74.8)	52.1 (38.0, 65.9)

1L first line, *2L* second line, *ALK* anaplastic lymphoma kinase, *CI* confidence interval, *EGFR* epidermal growth factor receptor, *IO* immuno-oncology, *PD-L1* programmed cell death ligand 1, *rwRR* real-world response rate

[a] Unweighted rwRR refers to analyses based on cohorts generated by only applying corresponding trial criteria, as feasible

[b] Weighted rwRR refers to the analyses performed on the real-world cohorts after inverse odds weights (based on published summary baseline characteristics of the trial populations) were applied in order to maximize the relevance of these cohorts to the intended comparison. Those were the cohorts used for the "weighted" analysis

[c] Weighted, confirmed rwRR refers to the analysis based on "confirmed responses" (as opposed to all responses) observed in the cohorts after weights were used

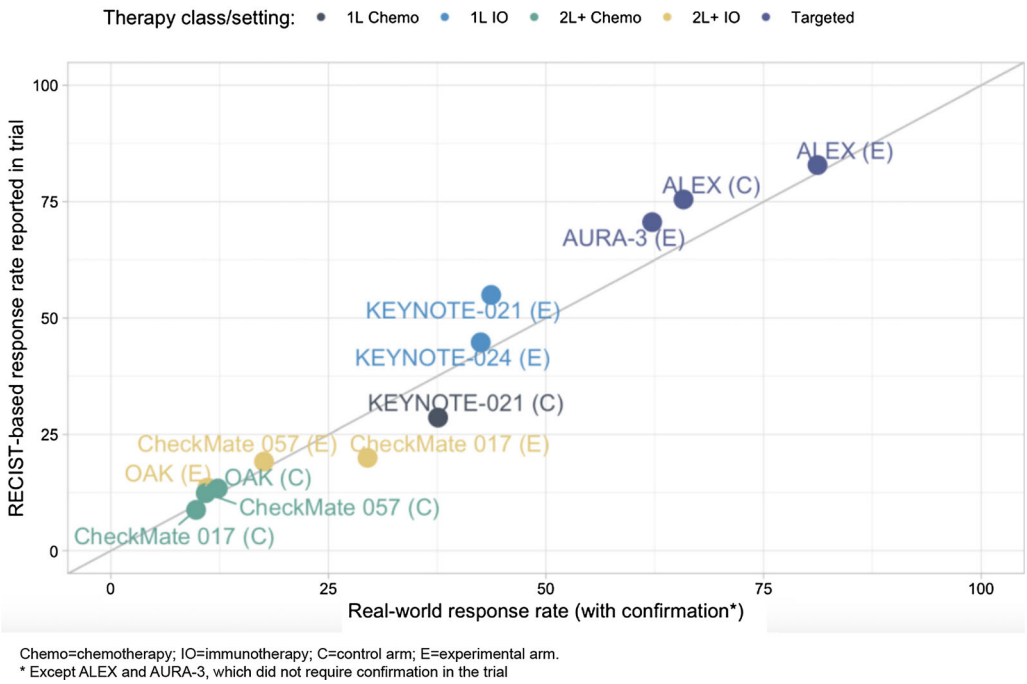

Fig. 3 Part 2 comparison and alignment of rwRR vs ORR in corresponding clinical trials

benchmarking exercise with results from the clinical trial endpoint showed that there may be a high correlation between rwRR and RECIST-based ORR. This finding is interesting, given that assessments during routine care differ from clinical trial assessments in several key points. First, real-world patient cohorts are likely to have a degree of heterogeneity far greater than clinical trial cohorts, not only in terms of patient and clinical characteristics but also in procedural aspects, such as the type and the timing intervals between assessments. In addition, routine assessments are traditionally considered more subjective and variable compared to RECIST, and may incorporate measures of a patient's overall condition beyond radiology [37]. Acknowledging these differences, our rwR variable, in concept, is not set to measure the same predefined elements as RECIST, but it seems to provide comparable information at the cohort level. These results were in line with a prior study to evaluate rwR-RECIST concordance at the patient level in metastatic breast cancer [38]

This study had limitations stemming from both the real-world data sources and from the methodology followed to assess the clinical relevance of rwR. Regarding the first issue, the integrity of our analyses and of the derivation of this variable relies on the availability of routine clinical information sources and their completeness in our EHR-derived database, which, in turn, are contingent on patterns of continuity and documentation in each originating clinic. This does not seem to have been problematic, since our results did not detect substantial data missingness relative to expectations set by professional practice guidelines [23]. Objectivity, or potential lack thereof, could also be a limitation of our sources, since these data depend on a manual abstraction process overlaid on top of the impressions reported by treating clinicians. We implemented abstraction procedures to strengthen reproducibility and robustness [9, 10], yet, the rates of inter-abstractor agreement indicate that approximately 1 in 5 assessments could be in question. As we point out earlier in the discussion, these agreement rates appear comparable to those observed during RECIST application; that context notwithstanding, the need to establish quality standards for RWE data

elements and variables remains open; such standards would enable researchers to determine whether this level of reproducibility is "good enough" to support wide adoption of this variable.

Related to the analyses performed to query the clinical relevance of this variable, some of the rules applied to optimize data analysis (exclusion of patients with 90-day data gaps, early progression, or death events) may have affected the correlations with rwPFS and rwOS. The trial benchmarking exercise included only trials that supported drug approvals by the US Food and Drug Administration (FDA), and we compared results at the cohort level, not at the patient level. Not every criterion of trial eligibility was matched to the real-world cohorts, and we cannot discount the potential effect of unmeasured confounders. Nonetheless, within the feasible alignment, the rwRR findings appeared to have strong correlation with the existing published results.

Altogether, these data demonstrate the utility of the rwR variable when confined to real-world outcomes research questions, such as comparative effectiveness (as measured by treatment response), or investigations of patient populations understudied in prospective trials. This response variable has already been included in a narrow-scope regulatory filing, where the use of RWD addressed a compelling clinical need (a rare population with limited treatment options) [39]. However, the utility of this variable in cases such as real-world comparator cohorts for single-arm clinical trials remains exploratory at this time. Additional research to study patient-level correlations with RECIST-based outcomes in cohorts with available radiographic images is required in order to consider more expansive uses; it is imperative to gain insight into the nature of this correlation, since it will have overarching influence driving alignment or misalignment between trial cohorts and real-world cohorts and will therefore be a defining factor in any comparison. Similarly, reaching a better understanding of the clinical meaning and context of response, progression, and treatment-based outcomes (often used as proxies for progression) in real-world cohorts will require exploring the correlations across achievement of response, rwPFS, and endpoints such as time-to-treatment discontinuation. Finding a response derivation approach with potential wide application across multiple tumor types and clinical settings will likely require additional adaptation and research, since different disease courses, clinical practices, and response kinetics may affect real-world documentation patterns. Furthermore, as RWD capture evolves, future research will likely reflect the increasing sophistication in routine care, e.g., incorporating newer imaging techniques or diagnostics such as circulating tumor markers.

CONCLUSIONS

We have developed a rwR variable derivation approach based on clinician assessments of disease burden documented in the EHR following radiographic evaluations. The resulting variable can provide a measure of treatment effectiveness. This study shows that EHR data curation can be leveraged to generate a feasible and reliable rwR variable, and associated endpoint analyses show that this variable is clinically meaningful. This type of research can help to fully realize the potential value of EHR-derived data. Future investigations are needed to evaluate patient-level concordance between real-world and RECIST-based endpoints and expand into other disease settings in order to understand the generalizability of these results across the oncology spectrum.

ACKNOWLEDGEMENTS

Funding. This study was sponsored by Flatiron Health, Inc., which is an independent subsidiary of the Roche group. Flatiron Health also funded the journal's Rapid Service and Open Access Fees.

Editorial Assistance. Julia Saiz, from Flatiron Health, for editorial support. Funded by Flatiron Health, Inc.

Authorship Contributions. Study design/concept: XM, LB, OT, CSB, MS, MT, NN, ABB. Data collection: Flatiron Health. Data interpretation and analysis: XM, LB, KM, CSB, OT, MS, MT, NN, ABB. Manuscript writing/review and approval: All.

Disclosures. All authors report employment at Flatiron Health, Inc., an independent subsidiary of the Roche group and stock ownership in Roche. LB, KM, OT, MS, MT, NN, BB, JK, ABB report equity ownership in Flatiron Health. The current affiliation for Caroline S Bennette is Institute for Disease Modeling, Gates Foundation, Seattle, WA, US.

Compliance with Ethics Guidelines. Institutional review board (IRB) approval of the study protocol, with a waiver of informed consent, was obtained prior to study conduct, and covers the data from all sites represented. Approval was granted by the WCG IRB. ("The Flatiron Health Real-World Evidence Parent Protocol", Tracking # FLI1-18-044).

Data Availability. The data that support the findings of this study have been originated by Flatiron Health, Inc. and are not publicly available, in order to safeguard the terms that ensure that the data remain deidentified. These deidentified data may be made available upon request, and are subject to a license agreement with Flatiron Health; interested researchers should contact DataAccess@flatiron.com to determine licensing terms.

REFERENCES

1. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-world evidence—what is it and what can it tell us? N Engl J Med. 2016;375(23):2293–7.

2. American Recovery and Reinvestment Act of 2009, Pub. L. No.111–5 (2009), Feb 17.

3. Franklin JM, Glynn RJ, Martin D, Schneeweiss S. Evaluating the use of nonrandomized real-world data analyses for regulatory decision making. Clin Pharmacol Ther. 2019;105(4):867–77.

4. US Food and Drug Administration. Guidance for industry: clinical trial endpoints for the approval of cancer drugs and biologics. Fed Regist. 2007;72.

5. Chen EY, Raghunathan V, Prasad V. An overview of cancer drugs approved by the US Food and Drug Administration based on the surrogate end point of response rate. JAMA Intern Med. 2019;179(7):915–21.

6. Chen EY, Haslam A, Prasad V. FDA acceptance of surrogate end points for cancer drug approval: 1992–2019. JAMA Intern Med. 2020;180(6):912.

7. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). Eur J Cancer. 2009;45(2):228–47.

8. Griffith SD, Tucker M, Bowser B, et al. Generating real-world tumor burden endpoints from electronic health record data: comparison of RECIST, radiology-anchored, and clinician-anchored approaches for abstracting real-world progression in non-small cell lung cancer. Adv Ther. 2019;36(8):2122–36.

9. Griffith SD, Miksad RA, Calkins G, et al. Characterizing the feasibility and performance of real-world tumor progression end points and their association with overall survival in a large advanced non-small-cell lung cancer data set. JCO Clin Cancer Inform. 2019;3:1–13.

10. Ma X, Long L, Moon S, Adamson BJS, Baxi SS. Comparison of population characteristics in real-world clinical oncology databases in the US:

Flatiron Health, SEER, and NPCR. medRxiv. 2020: 2020.03.16.20037143. http://medrxiv.org/content/early/2020/05/30/2020.03.16.20037143.abstract. https://doi.org/10.1101/2020.03.16.20037143.

11. Curtis MD, Griffith SD, Tucker M, et al. Development and validation of a high-quality composite real-world mortality endpoint. Health Serv Res. 2018;53(6):4460–76. https://doi.org/10.1111/1475-6773.12872.

12. Reck M, RodrÍguez-Abreu D, Robinson AG, et al. Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. N Engl J Med. 2016;375(19):1823–33.

13. Peters S, Camidge DR, Shaw AT, et al. Alectinib versus crizotinib in untreated ALK-positive non-small-cell lung cancer. N Engl J Med. 2017;377(9): 829–38.

14. Langer CJ, Gadgeel SM, Borghaei H, et al. Carboplatin and pemetrexed with or without pembrolizumab for advanced, non-squamous non-small-cell lung cancer: a randomised, phase 2 cohort of the open-label KEYNOTE-021 study. Lancet Oncol. 2016;17(11):1497–508.

15. Borghaei H, Paz-Ares L, Horn L, et al. Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. N Engl J Med. 2015;373(17): 1627–39.

16. Brahmer J, Reckamp KL, Baas P, et al. Nivolumab versus docetaxel in advanced squamous-cell non-small-cell lung cancer. N Engl J Med. 2015;373(2): 123–35.

17. Rittmeyer A, Barlesi F, Waterkamp D, et al. Atezolizumab versus docetaxel in patients with previously treated non-small-cell lung cancer (OAK): a phase 3, open-label, multicentre randomised controlled trial. Lancet. 2017;389(10066):255–65.

18. Mok TS, Wu Y, Ahn M, et al. Osimertinib or platinum-pemetrexed in *EGFR T790M*-positive lung cancer. N Engl J Med. 2017;376(7):629–40.

19. Westreich D, Edwards JK, Lesko CR, Stuart E, Cole SR. Transportability of trial results using inverse odds of sampling weights. Am J Epidemiol. 2017;186(8):1010–4.

20. Signorovitch JE, Sikirica V, Erder MH, et al. Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. Value Health. 2012;15(6):940–7.

21. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70(1):41–55.

22. Segal BD, Bennette CS. Transportability of trial results using inverse odds of sampling weights. Am J Epidemiol. 2018;187(12):2716–7.

23. Ettinger D, Wood ED, Aisner DL, et al. NCCN clinical practice guidelines in oncology non-small cell lung cancer (version 6.2020). https://www.nccn.org/professionals/physician_gls/pdf/nscl_blocks.pdf. Accessed 9 Oct 2020.

24. Suzuki C, Torkzad MR, Jacobsson H, et al. Interobserver and intraobserver variability in the response evaluation of cancer therapy according to RECIST and WHO-criteria. Acta Oncol. 2010;49(4):509–14.

25. Karmakar A, Kumtakar A, Sehgal H, Kumar S, Kalyanpur A. Interobserver variation in response evaluation criteria in solid tumors 1.1. Acad Radiol. 2019;26(4):489–501.

26. Ford RR, O'Neal M, Moskowitz SC, Fraunberger J. Adjudication rates between readers in blinded independent central review of oncology studies. J Clin Trials. 2016;6(5):289.

27. Feinberg BA, Bharmal M, Klink AJ, Nabhan C, Phatak H. Using response evaluation criteria in solid tumors in real-world evidence cancer research. Future Oncol. 2018;14(27):2841–8.

28. Luke JJ, Ghate SR, Kish J, et al. Targeted agents or immuno-oncology therapies as first-line therapy for BRAF-mutated metastatic melanoma: a real-world study. Future Oncol. 2019;15(25):2933–42.

29. Kehl KL, Elmarakeby H, Nishino M, et al. Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. JAMA Oncol. 2019;5(10):1421–9.

30. Stewart M, Norden AD, Dreyer N, et al. An exploratory analysis of real-world end points for assessing outcomes among immunotherapy-treated patients with advanced non-small-cell lung cancer. JCO Clin Cancer Inform. 2019;3:1–15.

31. Arbour KC, Luu AT, Luo J, et al. Deep learning to estimate RECIST in patients with NSCLC treated with PD-1 blockade. Cancer Discov. 2020;5:6. https://doi.org/10.1158/2159-8290.CD-20-0419.

32. O'Connell JP, Kris MG, Gralla RJ, et al. Frequency and prognostic importance of pretreatment clinical characteristics in patients with advanced non-small-cell lung cancer treated with combination chemotherapy. J Clin Oncol. 1986;4(11):1604–14.

33. Sørensen JB, Badsberg JH, Hansen HH. Response to cytostatic treatment in inoperable adenocarcinoma of the lung: critical implications. Br J Cancer. 1989;60(3):389–93.

34. Paesmans M, Sculier JP, Libert P, et al. Response to chemotherapy has predictive value for further survival of patients with advanced non-small cell lung cancer: 10 years experience of the European Lung Cancer Working Party. Eur J Cancer. 1997;33(14): 2326–32.

35. Akerley W, Crowley J, Giroux D, Gandara D. Response to chemotherapy as a predictor of survival in advanced non-small cell lung cancer (NSCLC): review of the Southwest Oncology Group (SWOG) database. Lung Cancer. 2000;29(1):33.

36. Blumenthal GM, Karuri SW, Zhang H, et al. Overall response rate, progression-free survival, and overall survival with targeted and standard therapies in advanced non-small-cell lung cancer: US Food and Drug Administration trial-level and patient-level analyses. J Clin Oncol. 2015;33(9):1008.

37. Therasse P. Measuring the clinical response. What does it mean? Eur J Cancer. 2002;38(14):1817–23.

38. Huang Bartlett C, Mardekian J, Cotter MJ, et al. Concordance of real-world versus conventional progression-free survival from a phase 3 trial of endocrine therapy as first-line treatment for metastatic breast cancer. PLoS One. 2020;15(4): e0227256.

39. Wedam S, Fashoyin-Aje L, Bloomquist E, et al. FDA approval summary: palbociclib for male patients with metastatic breast cancer. Clin Cancer Res. 2020;26(6):1208–12.