Replication of Overall Survival, Progression-Free Survival, and Overall Response in Chemotherapy Arms of Non–Small Cell Lung Cancer Trials Using Real-World Data

Thanh G.N. Ton[1], Navdeep Pal[1], Huong Trinh[1], Sami Mahrus[1], Michael T. Bretscher[2], Robson J.M. Machado[3], Natalia Sadetsky[1], Nayan Chaudhary[1], Michael W. Lu[1], and Gregory J. Riely[4]

ABSTRACT

Purpose: The utility of real-world data (RWD) for use as external controls in drug development is informed by studies that replicate trial control arms for different endpoints. The purpose of this study was to replicate control arms from four non–small cell lung cancer (NSCLC) randomized controlled trials (RCT) to analyze overall survival (OS), progression-free survival (PFS), and overall response rate (ORR) using RWD.

Patients and Methods: This study used RWD from a nationwide de-identified database and a clinico-genomic database to replicate OS, PFS, and ORR endpoints in the chemotherapy control arms of four first-line NSCLC RCTs evaluating atezolizumab [IMpower150–wild-type (WT), IMpower130-WT, IMpower131, and IMpower132]. Additional objectives were to develop a definition of real-world PFS (rwPFS) and to evaluate the real-world response rate (rwRR) endpoint.

Results: Baseline demographic and clinical characteristics were balanced after application of propensity score weighting methods. For rwPFS and OS, RWD external controls were generally similar to their RCT control counterparts. Across all four trials, the hazard ratio (HR) point estimates comparing trial controls with external controls were closer to 1.0 for the PFS endpoint than for the OS endpoint. An exploratory assessment of rwRR in RWD revealed a slight but nonsignificant overestimation of RCT ORR, which was unconfounded by baseline characteristics.

Conclusions: RWD can be used to reasonably replicate the OS and PFS of chemotherapy control arms of first-line NSCLC RCTs. Additional studies can provide greater insight into the utility of RWD in drug development.

Introduction

With the recent and rapid adoption of electronic health records (EHR) to document patient information at the point of care, real-world data (RWD) have become a viable source of clinical information (1). Since the passing of the 21st Century Cures Act into law to accelerate medical development and bring innovations faster and more efficiently to patients (2), greater focus has emphasized the use of real-world evidence to help inform treatment effectiveness and regulatory decisions (3, 4). In certain instances when randomization to a control arm is either unethical or infeasible, RWD have supported drug approval (3) or label expansion (5) decisions. In single-arm trial designs, RWD can provide comparative benchmarks, particularly for treatment of rare diseases with high unmet needs (1) or those with breakthrough designations (6). For RWD external controls to meaningfully inform internal or regulatory decision-making, not only must the RWD source have high standards for data reliability and relevance, but it must also be accompanied by analytic approaches that ensure selection of comparable populations, similar outcome measurements, and appropriate statistical methods (3).

The ability of RWD to serve as external controls to single-arm trials was explored in a study in which external controls derived from EHR data replicated overall survival (OS) for the control arms of eight randomized controlled trials (RCT) among patients with non–small cell lung cancer (NSCLC; ref. 7). However, the primary endpoints of most oncology studies leading to regulatory approval for uncommon subsets of cancers are often overall response rate (ORR) or progression-free survival (PFS) endpoints. Therefore, exploration of RWD endpoints other than OS can add valuable insights and potentially expand applications of RWD in drug development.

The primary objectives of this study were to replicate OS in the intent-to-treat population of patients who were randomized to receive chemotherapy in first-line NSCLC trials evaluating atezolizumab that, at the time, had an interim analysis for both PFS and OS endpoints. The following trials were included: (i) IMpower150–wild-type (WT) patient population that does not harbor epidermal growth factor receptor (*EGFR*)/anaplastic lymphoma kinase (*ALK*) mutations (NCT02366143; ref. 8); (ii) IMpower130-WT population without *EGFR*/*ALK* mutations (NCT02367781; ref. 9); (iii) IMpower131 population with *EGFR* mutations or *ALK* rearrangements if patients had progressed on or were intolerant of tyrosine kinase inhibitors (NCT02367794; ref. 10); and (iv) IMpower132 population without *EGFR*/*ALK* mutations (NCT02657434; ref. 11). The second objective was to explore a feasible definition of the real-world PFS (rwPFS) endpoint using EHR data as well as replicate PFS in chemotherapy control arms in the same four IMpower trials. Finally, the last objective was to explore the real-world response rate (rwRR) endpoint and to replicate ORR in control arms in these same IMpower trials.

[1]Genentech, Inc., South San Francisco, California. [2]F. Hoffmann-La Roche Ltd., Basel, Switzerland. [3]Roche Products, Welwyn Garden City, UK. [4]Memorial Sloan Kettering Cancer Center, New York, New York.

Note: Supplementary data for this article are available at Clinical Cancer Research Online (http://clincancerres.aacrjournals.org/).

Corresponding Author: Thanh G.N. Ton, Genentech, Inc., 1 DNA Way, South San Francisco, CA 94080. Phone: 206-375-9710; E-mail: tont2@gene.com

Clin Cancer Res 2022;28:2844–53

doi: 10.1158/1078-0432.CCR-22-0471

Translational Relevance

Exploration of real-world data (RWD) endpoints other than overall survival (OS) can provide valuable insights into expanding applications of RWD in drug development. This analysis explored several large RWD sets and clinical trials to develop approaches to determine whether real-world endpoints are similar to the progression-free survival (PFS), OS, and overall response rate (ORR) endpoints used in clinical trials. Applying population restriction and propensity score methods to replicate trial control arms with RWD external controls performed reasonably well for real-world PFS and OS for four selected non–small cell lung cancer trials of atezolizumab given in the first-line setting. General concordance of real-world response rates with ORR from trial data was observed. Extending this work to different treatments, tumor types, lines of therapy, and a thorough exploration of differences in scan frequencies in various settings will help provide greater insight into the utility of RWD in drug development.

Patients and Methods

Data sources

For each RCT, patient-level data represented at the time of interim analysis were used. Only patients randomized to the chemotherapy control arms in each trial were included. For RWD, the nationwide Flatiron Health (FH) EHR-derived de-identified database was used for OS and rwPFS endpoints. The FH database is a longitudinal database, comprising de-identified patient-level structured and unstructured data, curated via technology-enabled abstraction (12, 13). The de-identified data originated from approximately 280 US cancer clinics (~800 sites of care) and included patients diagnosed with advanced or metastatic NSCLC on or after January 1, 2011. The majority of patients in the database received care in the community oncology setting. The starting size of the Flatiron database used for this study was 60,517. Institutional Review Board approval of the study protocol was obtained prior to study conduct and included a waiver of informed consent, and studies were conducted in accordance with recognized ethical guidelines (e.g., Declaration of Helsinki, CIOMS, Belmont Report, U.S. Common Rule).

For evaluation of the rwRR endpoint, the nationwide de-identified FH–Foundation Medicine Inc (FMI) NSCLC clinico-genomic database (CGDB) was used because rwRR was abstracted and available only in the CGDB (rwRR was not abstracted in the FH EHR data). Retrospective longitudinal clinical data were derived from EHR data, comprising patient-level structured and unstructured data curated via technology-enabled abstraction, and were linked to genomic data derived from FMI comprehensive genomic profiling tests in the FH-FMI CGDB by de-identified, deterministic matching (14). The CGDB includes patients diagnosed with NSCLC on or after January 1, 2011, who underwent FMI comprehensive genomic profile testing, although next-generation sequencing test results were not used in this analysis. Information specifically on *EGFR*/*ALK* was available as part of the abstraction of medical records by FH. The specific data set cuts used in each analysis were dependent on the availability of the abstracted rwRR variable. While the CGDB does contain information on OS and rwPFS endpoints, the CGDB was not used for the OS and rwPFS endpoints given that its sample size is a much smaller starting sample size compared with the FH EHR data source.

Study population

Patients who met the eligibility criteria were identified from the EHR data. Cohort attrition was specific to each IMpower trial for first-line chemotherapy regimens, histologies, and *EGFR*/*ALK* inclusion/exclusion criteria (Supplementary Fig. S1). Index date was defined as the date of first-line treatment initiation in the RWD and the date of randomization in the corresponding trial. Baseline Eastern Cooperative Oncology Group performance status (ECOG PS) values were defined within −30 to +7 days of index date, and laboratory test results were defined within −28 to 0 days of index date. Baseline *EGFR*/*ALK* status was defined as results from tests any time before and up to +7 days of index date. Patients with missing *EGFR*/*ALK* test results were included in the primary analysis and later excluded as part of sensitivity analyses. Patients with abnormal hematologic and organ function were excluded. Adequate hematologic and organ function was defined as having normal values for all of the following laboratory measures: absolute neutrophil count ≥1,500 cells/μL, absolute lymphocyte count ≥500/μL, white blood cell count ≥2,500/μL, serum albumin ≥2.5 g/dL, platelet count ≥100,000/μL, hemoglobin ≥9.0 g/dL, aspartate aminotransferase ≤2.5 × upper limit of normal (ULN), alanine aminotransferase ≤2.5 × ULN, alkaline phosphatase ≤2.5 × ULN, serum bilirubin ≤1.0 × ULN, serum creatinine ≤1.5 × ULN, and serum calcium ≤12.0 mg/dL. Patients with missing laboratory results were not excluded from the cohort.

For the rwRR endpoint, patients were identified from the CGDB. Due to the exploratory nature of this endpoint and the smaller starting sample size of the CGDB, analyses were conducted incrementally according to progressively more strict cohort attrition groups (Supplementary Table S1) in order to evaluate the impact of population restriction on sample size. Restriction criteria Group 4 represents the analytic cohort that most closely resembles the full set of criteria used to assemble the final trial-like cohorts for OS and rwPFS using the EHR data.

Endpoints

The mortality endpoint was previously developed as a composite variable linking multiple internal and external data sources. A validation study demonstrated a sensitivity of 89.7% and specificity of 97.3% for the mortality variable with a positive predictive value of nearly 98% when validated against the national death index (15, 16).

Real-world progression was abstracted retrospectively from information documented in the EHRs as part of routine clinical practice to indicate when a patient's cancer progressed. Abstractors reviewed clinician notes to identify evidence of progression, which may have been documented directly (e.g., "the patient has progressive disease") or indirectly (e.g., "the patient's cancer burden is worse"), or acknowledged by source evidence consistent with progression (e.g., "radiology report that finds new lesions or increased size of existing lesions"). For each evidence of a progression event, abstractors noted the date and the type of source evidence (e.g., radiology scan, pathology report via biopsy). Progression events are distinct episodes in the patient journey at which the treating clinician concludes that there has been spread or worsening of disease, and are updated every 6 months within the EHR data.

The rwPFS endpoint is a composite variable defined by: (i) inclusion of all progression events from all sources of evidence (not only radiographically confirmed); (ii) inclusion of all progression events, even if occurrence was within the initial 14 days of index date; (iii) not defining line of therapy (LOT) advancement as a progression event; (iv) inclusion of deaths as rwPFS events within a 30-day window after the end of progression follow-up; and (v) censoring at the beginning of visit gaps >90 days. Exploration of alternative rwPFS definitions included omission of progression events without radiographic confirmation or events immediately after baseline (<14 days), censoring at LOT change, and varying the time window for inclusion of death events after the end of progression follow-up (0, 10, 30, 60, 90, or 180 days). The effects on HRs and median survival compared with the main definition were minor across all evaluations.

rwRR was extracted retrospectively from EHRs of patients in the CGDB for selected treatments. The process of abstraction has been described in detail elsewhere (17). In brief, abstractors documented clinicians' assessments or interpretations of radiologic scans for change in disease burden in an individual patient during a LOT, which may be acknowledged directly (e.g., "positive interval response to treatment") or indirectly (e.g., "adenopathy is stable"). Scans of any modality, which fall within 30 days of treatment initiation were not evaluated for response in order to provide sufficient time for treatment effect. Abstractors bundled scans within a 14-day period and documented the assessment date as the earliest date of each 14-day period. Possible categories included real-world complete response (CR), real-world partial response (PR), real-world stable disease, real-world progressive disease, real-world pseudoprogression, indeterminate response, and not documented. For these analyses, real-world response was dichotomized as CR or PR. Nonresponders included patients with stable or progressive disease, indeterminate response, and not documented.

Statistical methods

Analytic framework for OS and rwPFS

The analysis of IMpower150-WT was used to optimize and finalize the analytic framework for the other three trial comparisons, including decisions regarding primary versus sensitivity analyses, management of missing biomarker data for *EGFR*/*ALK*, covariates for propensity score models, and weighting and trimming choices. For OS, the analytic approach relied heavily on work by Carrigan and colleagues (7) in NSCLC because IMpower150 was one of the many trials replicated in this previously published work. For rwPFS, exploratory analyses were conducted using IMpower150-WT to characterize each component of the composite endpoint. Once finalized, the analytic framework became *a priori* methods for replication of control arms of the remaining three trials.

In the final analytic framework, OS was defined as time from randomization (RCT) or first-line start (RWD external control) to death, last patient visit (RWD external control), or end of follow-up (RCT). rwPFS was defined as time from randomization (RCT) or first-line start (RWD external control) to first progression event or death. Patients without progression or death were censored at the beginning of a >90-day visit gap or last clinic note, whichever event occurred first. For each trial, and separately for each endpoint, we sought to: (i) replicate trial control arms by comparing the RWD external control versus the RCT control (reference), and (ii) replicate the treatment effect estimate of each clinical trial by replacing the trial control arm with RWD external control (reference). Propensity score models included the same prognostic variables for all endpoints: age, race, sex, stage at diagnosis, smoking, and time from initial diagnosis to first-line start date. Differences in baseline characteristics between RWD external control and RCT control were balanced using inverse probability weights, for which only patients in the RWD external control were re-weighted and trimmed, when necessary, corresponding to the estimand for average treatment effect on the treated (ATT). This weighting framework is also referred to as standardized mortality ratio weighting (18). Inverse probability treatment weighting (IPTW) was chosen over other methods (e.g., PS stratification, PS 1:1 matching, PS covariate adjustment) because it was demonstrated by Carrigan and colleagues that results did not substantively differ across different PS methods (7). For each covariate, balance between the two arms was evaluated using standard mean difference (SMD), for which ideal balance was defined as SMD <0.1. Weighted Kaplan–Meier curves were used to compare the two arms, and Cox regression models were used to obtain HRs and 95% confidence intervals (CI). The starting sample size of the weighted pseudo-population is provided in Supplementary Table S2.

Exploratory analytic framework for rwRR

Due to the small starting sample size of the CGDB used for rwRR and the exploratory nature of this endpoint, rwRR was calculated for each restriction criteria group (Supplementary Table S1) to evaluate the impact of population restriction on RRs as well as to closely monitor the decline in sample size. Unconfirmed ORR from RCTs were compared with unconfirmed rwRR in RWD external controls. Crude RRs were calculated as the proportion of patients who had evidence of either PR or CR among all patients identified within each restriction cohort. Patients who had missing response data were included in the denominator. Probability weights were derived from logistic models that included age, race, sex, stage at diagnosis, smoking, and time from initial diagnosis to first-line start date. Weighted rwRR and corresponding 95% CIs were reported. A sensitivity analysis was conducted in which patients with missing *EGFR* or *ALK* mutation test results were excluded.

Data availability

Qualified researchers may request access to individual patient-level data through the clinical study data request platform (https://vivli.org/). Further details on Roche's criteria for eligible studies are available here (https://vivli.org/members/ourmembers/). For further details on Roche's Global Policy on the Sharing of Clinical Information and how to request access to related clinical study documents, see here: https://www.roche.com/research_and_development/who_we_are_how_we_work/clinical_trials/our_commitment_to_data_sharing.htm.

The data that support the findings of this study have been originated by Flatiron Health, Inc. and Foundation Medicine, Inc. These deidentified data may be made available upon request and are subject to a license agreement with Flatiron Health and Foundation Medicine; interested researchers should contact cgdb-fmi@flatiron.com and dataaccess@flatiron.com to determine licensing terms.

Results

rwPFS and OS

For deriving a trial-like population, as many inclusion and exclusion criteria were implemented as possible to emulate each trial as closely as possible using RWD (age, histology, presence of driver mutations, ECOG PS, treatment regimen, no prior cancer therapy, and normal laboratory measures). Several additional criteria were implemented for the RWD to further minimize potential misclassification (confirmed receipt of treatment with administration data; exclusion of patients with >90-day visit gaps to address potential migration out of the EHR system; Supplementary Fig. S1). The final sample sizes for the RWD external cohorts obtained from EHRs were $n = 436$ for IMpower150-WT; $n = 118$ for IMpower130-WT; $n = 493$ for IMpower131; and $n = 2{,}011$ for IMpower132. In general, patients in the RWD external cohorts were

Table 1. Baseline demographics for patients in the IMpower RCTs and RWD cohorts for PFS and OS in the Flatiron Health de-identified EHR-derived NSCLC database and real-world response in the de-identified Flatiron Health–FMI clinico-genomic NSCLC database.

	OS and PFS								Real-world response[a]			
	IMpower150-WT		IMpower130-WT		IMpower131		IMpower132					
Categories	**RCT (*n* = 338)**	**RWD (*n* = 436)**	**RCT (*n* = 228)**	**RWD (*n* = 118)**	**RCT (*n* = 340)**	**RWD (*n* = 493)**	**RCT (*n* = 286)**	**RWD (*n* = 2,011)**	**IMpower 150-WT RWD (*n* = 61)**	**IMpower 130-WT RWD (*n* = 15)**	**IMpower 131 RWD (*n* = 57)**	**IMpower 132 RWD (*n* = 281)**
Male, *n* (%)	209 (61.8)	228 (52.3)	134 (58.8)	61 (51.7)	277 (81.5)	331 (67.1)	192 (67.1)	1,015 (50.5)	30 (49.2)	8 (53.3)	38 (66.7)	137 (48.8)
White, *n* (%)	298 (88.2)	254 (58.3)	210 (92.1)	89 (75.4)	290 (85.3)	348 (70.6)	203 (71.0)	1,437 (71.5)	36 (59.0)	14 (93.3)	44 (77.2)	201 (71.5)
Age at randomization/first-line start, *n* (%)												
<65 y	204 (60.4)	227 (52.1)	120 (52.6)	41 (34.7)	177 (52.1)	159 (32.3)	173 (60.5)	869 (43.2)	31 (50.8)	8 (53.3)	17 (29.8)	124 (44.1)
65–75 y	107 (31.7)	144 (33.0)	86 (37.7)	37 (31.4)	132 (38.8)	207 (42.0)	97 (33.9)	698 (34.7)	21 (34.4)	<5	26 (45.6)	89 (31.7)
≥75 y	27 (8.0)	65 (14.9)	22 (9.6)	40 (33.9)	31 (9.1)	127 (25.8)	16 (5.6)	444 (22.1)	9 (14.8)	<5	14 (24.6)	68 (24.2)
Duration from initial diagnosis to index date, median [IQR], months	1.66 [1.08–2.71]	1.17 [0.76–2.04]	1.58 [1.08–2.99]	1.35 [0.83–4.02]	1.38 [0.91–3.76]	1.18 [0.72–2.53]	1.35 [0.92–2.48]	1.31 [0.82–2.43]	1.05 [0.69–1.61]	1.15 [0.72–1.61]	1.12 [0.75–1.91]	1.28 [0.89–2.14]
Stage at diagnosis, *n* (%)												
I	17 (5.0)	25 (5.7)	19 (8.3)	18 (15.3)	18 (5.3)	44 (8.9)	7 (2.4)	157 (7.8)	<5	0 (0.0)	<5	27 (9.6)
II	15 (4.4)	9 (2.1)	11 (4.8)	4 (3.4)	24 (7.1)	21 (4.3)	11 (3.8)	74 (3.7)	<5	0 (0.0)	<5	<5
III	29 (8.6)	45 (10.3)	14 (6.1)	14 (11.9)	41 (12.1)	97 (19.7)	21 (7.3)	337 (16.8)	8 (13.1)	0 (0.0)	7 (12.3)	36 (12.8)
IV	267 (79.0)	353 (81.0)	183 (80.3)	81 (68.6)	254 (74.7)	323 (65.5)	246 (86.0)	1,423 (70.8)	51 (83.6)	15 (100.0)	43 (75.4)	213 (75.8)
Other	10 (3.0)	<5	<5	<5	<5	<5	<5	20 (1.0)	<5	<5	<5	<5
History of smoking, *n* (%)	288 (85.2)	390 (89.4)	211 (92.5)	107 (90.7)	316 (92.9)	477 (96.8)	256 (89.5)	1,802 (89.6)	56 (91.8)	15 (100.0)	52 (91.2)	241 (85.8)

Abbreviation: IQR, interquartile range.

[a] RCT demographics for the response assessment were the same used for the OS and PFS assessments.

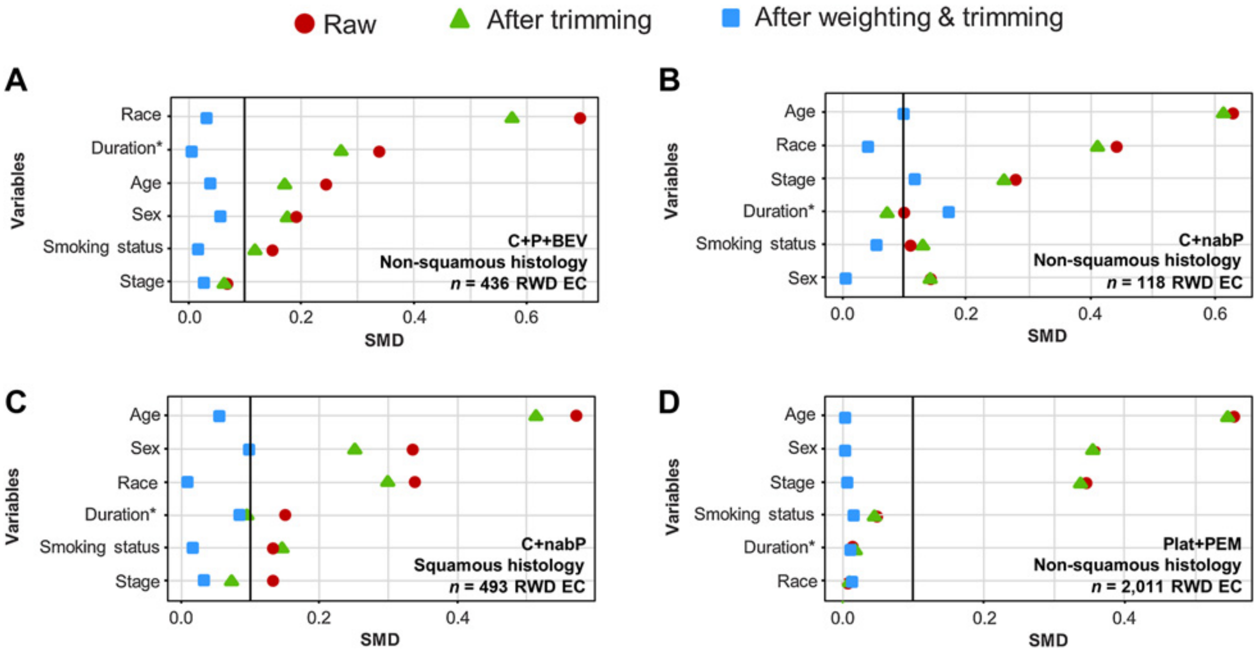

Figure 1.

Adjustment of baseline characteristics between the RCT and RWD arms for OS and rwPFS assessments. Adjustments of characteristics for IMpower150-WT (**A**), IMpower130-WT (**B**), IMpower131 (**C**), and IMpower132 (**D**) where red indicates raw data, green indicates data with trimming, and blue indicates data after weighting and trimming. Standardized mortality ratio with no trimming was applied to patients in the RCT. For final analysis, inverse probability weighting with average treatment effect for the treated population estimates with no trimming were applied to the RCT cohort to achieve cohort balancing. Ideal balance occurs when the data have SMD < 0.1 (bold line) for all baseline characteristics. BEV, bevacizumab; C, carboplatin; EC, external control; nabP; nab-paclitaxel; P, paclitaxel; PEM, pemetrexed; Plat, carboplatin/cisplatin. *, Duration was defined as the time from initial diagnosis to index date.

older, more likely to be female, less likely to be White, and diagnosed with *de novo* metastatic disease (**Table 1**).

After weighting and trimming, baseline demographics and clinical characteristics were balanced between the RCT control and RWD external control arms (SMD <0.1), with the exception of IMpower130-WT, in which the median time from initial diagnosis to index date was shorter in the RWD external control arm compared with the trial control arm (**Fig. 1**). For rwPFS, the proportion of death events within the composite endpoint (death and progression combined) was higher in the RWD external control compared with the corresponding trial for IMpower130-WT (24.4% vs. 15.1%) and IMpower131 (24.1% vs. 18.5%) but was similar for IMpower150-WT (15.6% vs. 17.5%) and IMpower132 (17.7% vs. 15.1%).

Kaplan–Meier curves showed the replication of control arms using RWD. For the rwPFS (**Fig. 2**) and OS (**Fig. 3**) endpoints, RWD external control arms were similar to their RCT control counterparts, with the exception of the OS endpoint for IMpower130-WT and IMpower131, for which the RWD external control showed decreased survival compared with the RCT control. HR estimates for replicating RCT controls were summarized for PFS (**Fig. 4A**) and OS (**Fig. 4B**). For these analyses, the RCT control was the reference, such that HR = 1.0 represented perfect replication of the trial control using the RWD external control. An HR >1 suggested that the RWD external control arm had a worse outcome, while an HR <1 suggested that the RWD external control arm had a better outcome. The HR point estimates across all four trials were closer to 1.0 for the PFS endpoint compared with the OS endpoint, with notably worse performance in the RWD external controls for IMpower131 and IMpower130-WT for the OS endpoint.

Results for rwRRs

For rwRR, the sample sizes of the most restrictive group in the RWD arm represented in Group 4 were $n = 61$ patients for IMpower150-WT; $n = 15$ for IMpower130-WT; $n = 57$ for IMpower131; and $n = 281$ for IMpower132 (Supplementary Table S1). Except for IMpower130-WT, for which the RWD external control sample size was too small for interpretable comparisons, patients in the RWD cohorts were less likely to be men and were typically older compared with those in the corresponding RCT control arms (**Table 1**). Time from initial diagnosis to index date was generally shorter for patients in the RWD external control arms. The distribution of stage at initial diagnosis varied between RWD and RCT controls across different trials. Unadjusted point estimates of rwRR in the RWD external control replicating IMpower150-WT bevacizumab + carboplatin + paclitaxel were approximately 10 to 15 percentage points higher (60.7%–65.3%) compared with the corresponding ORR (48.8%) in the RCT, regardless of how the RWD population was restricted (**Fig. 5**). For IMpower130-WT, rwRR estimates were 15 to 20 percentage points higher (54.2%–60.0%) than those observed for the RCT, although CIs were extremely wide due to the small sample sizes. The difference in point estimates between the RWD external control and RCT control was <10 percentage points for IMpower132. RCT ORR and rwRR were most similar for the IMpower131 RCT. CIs around point estimates for weighted rwRR, unweighted rwRR, and RCT ORR were all overlapping (Supplementary Table S1). Of note, weighting response rates in the RWD external controls to match baseline characteristics of corresponding RCT controls did not appreciably alter the unweighted estimates (Supplementary Table S3). In closer assessment of individual response categories, the trend in overestimation of rwRR relative to

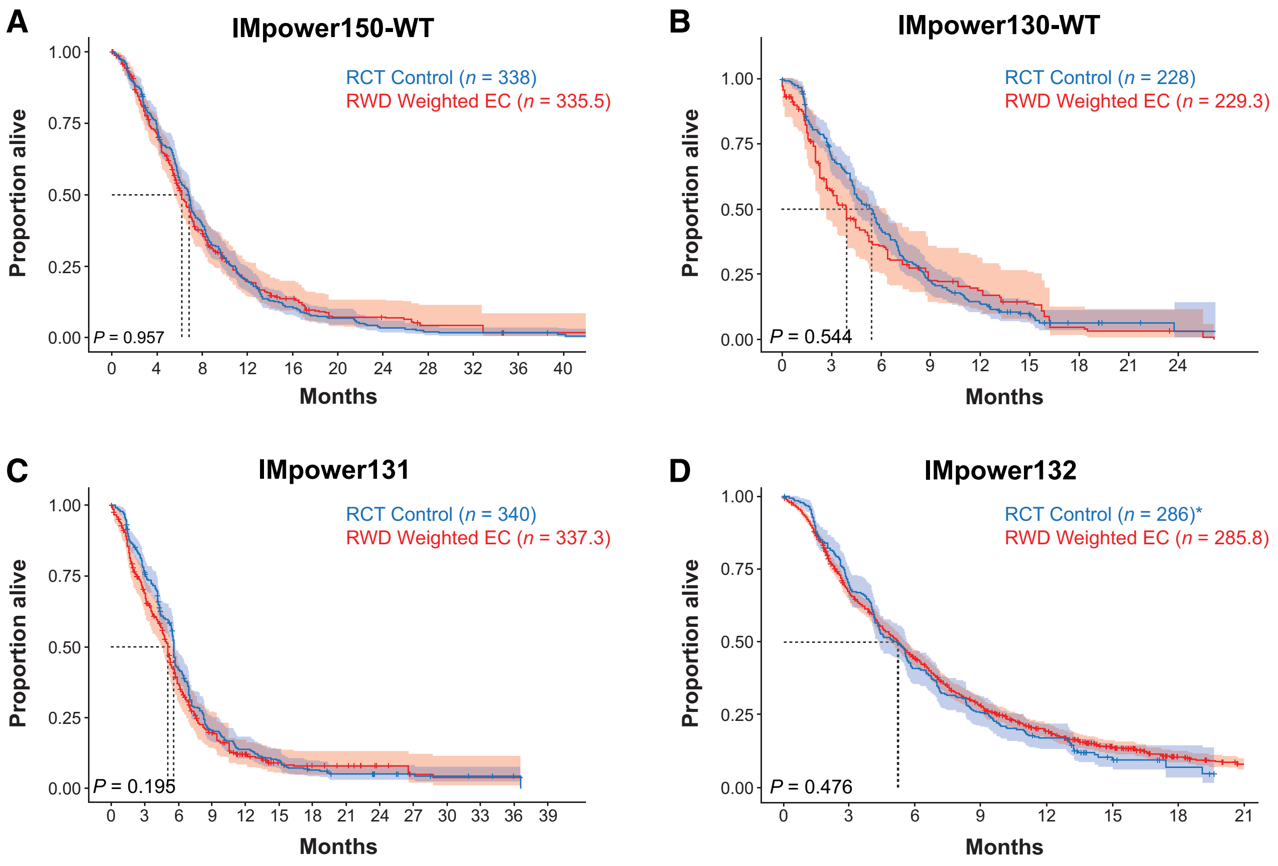

Figure 2.

Primary results to replicate PFS of RCT control arms using RWD. Kaplan–Meier curves comparing PFS for the RCT control arms versus the RWD EC arms in the IMpower150-WT (**A**), IMpower130-WT (**B**), IMpower131 (**C**), and IMpower132 RCTs (**D**). Standard mortality ratio weighting without trimming was performed. Sample sizes may differ from OS, as some patients lacked follow-up data extending beyond baseline. EC, external control. *, WT intent-to-treat population without epidermal growth factor/anaplastic lymphoma kinase alterations.

RCT ORR may have been due to higher numbers of PRs in the RWD arm (Supplementary Fig. S2), although no statistical significance testing or 95% CIs were generated for this more granular comparison.

Discussion

For endpoints other than OS, especially those that typically depend on radiographic criteria such as for PFS and ORR, the utility of using RWD to serve as external controls is currently unclear. Our current study builds upon the current literature by: (i) extending the work of Carrigan and colleagues for the OS endpoint to include more recent IMpower trials; (ii) replicating trial control arms with rwPFS and rwRR endpoints for the first time in these IMpower trials; and (iii) improving upon the methodology used in recent publications by incorporating patient-level data for clinical trials (7, 19).

In general, applying population restriction and propensity score methods to replicate RCT control arms with RWD external controls performed reasonably well for rwPFS and OS for the four selected IMpower trials, with notably better replication for the rwPFS endpoint compared with the OS endpoint. It was not surprising that treatment regimens that are uncommon in routine clinical settings made balancing between arms difficult, such as in the case of IMpower130-WT (which used nab-paclitaxel for treatment of patients with non-squamous NSCLC, for which carboplatin + pemetrexed is over 10 times more commonly used in real-world settings). As a result of the small sample size, the RWD external control had a lower median time from initial diagnosis to index date, which potentially contributed to worse survival compared with the IMpower130-WT control (20). In general, these results were consistent with previous attempts at replicating OS in the control arms of eight different trials in NSCLC (7).

Considerations for the rwPFS endpoint are much more complex because identifying progression events in the EHR using the RECIST approach is less feasible than the clinician-anchored approach supported by radiology report data (21). rwPFS abstracted from EHR data across diverse healthcare data organizations has demonstrated clinically relevant correlations with other intermediate real-world endpoints and can produce findings that are directionally similar to those from trials, despite measuring progression differently in real-world settings (20). Better replication was observed with the rwPFS endpoint compared with the OS endpoint, consistent with findings from Tan and colleagues, and likely explained by the greater impact of post-baseline confounding factors on OS results than PFS results (19). These results were also consistent with a similar study in metastatic breast cancer, in which the use of EHR-derived data and similar IPTW methods resulted in similar median estimates for rwPFS and

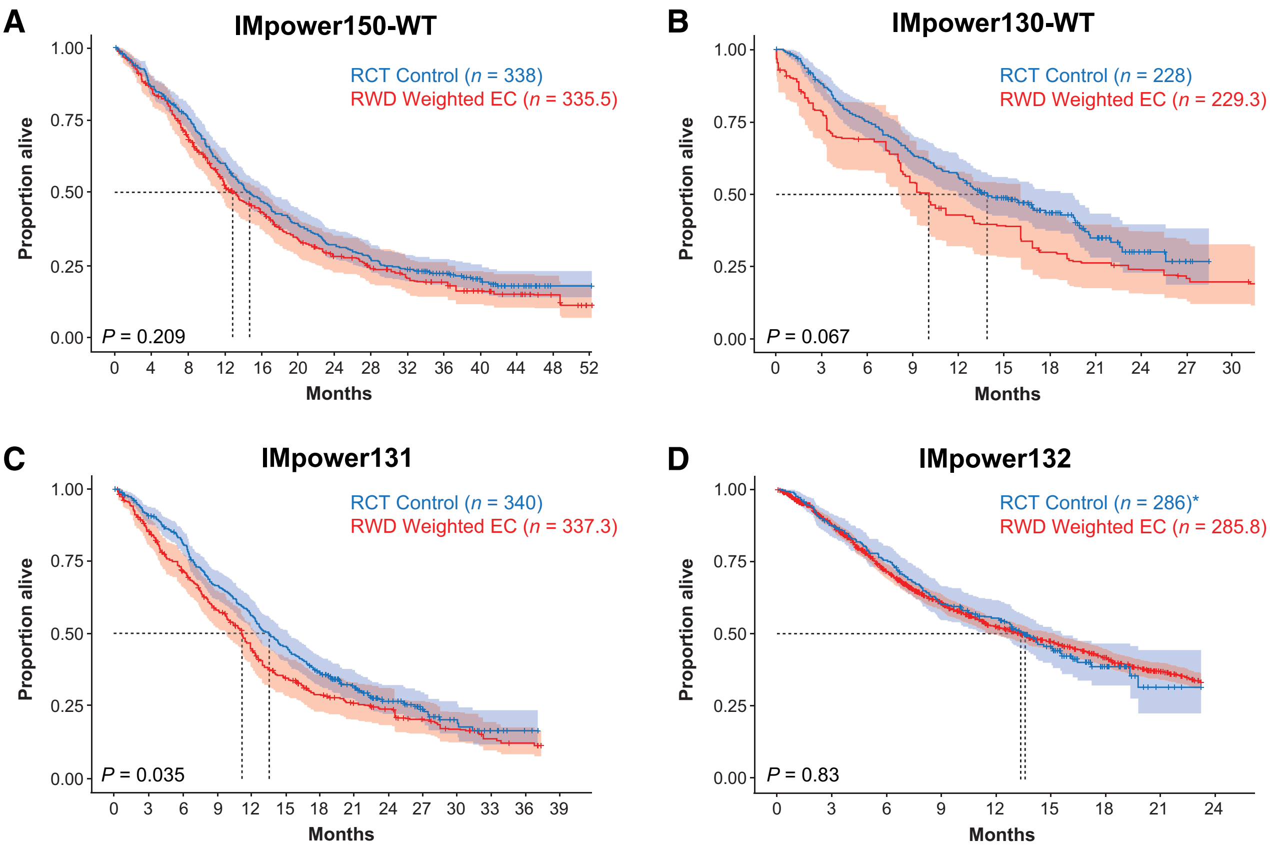

Figure 3.

Primary results to replicate OS of RCT control arms using RWD. Kaplan–Meier curves comparing OS for the RCT control arms versus the RWD EC arms in the IMpower150-WT (**A**), IMpower130-WT (**B**), IMpower131 (**C**), and IMpower132 RCTs (**D**). Standard mortality ratio weighting without trimming was performed. Sample sizes may differ from PFS, as some patients lacked follow-up data extending beyond baseline. EC, external control. *, WT intent-to-treat population without epidermal growth factor/anaplastic lymphoma kinase alterations.

RECIST-PFS (18.4 vs. 16.6 months) in the letrozole control arm of the PALOMA-2 trial (weighted HR = 1.04; 95% CI, 0.69–1.56; ref. 22).

While we saw general concordance in rwRR with the ORR RCT data, several observations were notable. First, when point estimates of response rates differed between the RWD and RCTs, rwRRs in RWD were higher than the ORR from RCTs (although 95% CIs were overlapping). Because the rwRR derived here was not based on measurement of images but rather interpretations of images, small reductions in tumors that would not meet RECIST criteria may have been categorized as responses. An overestimation of rwRR has been

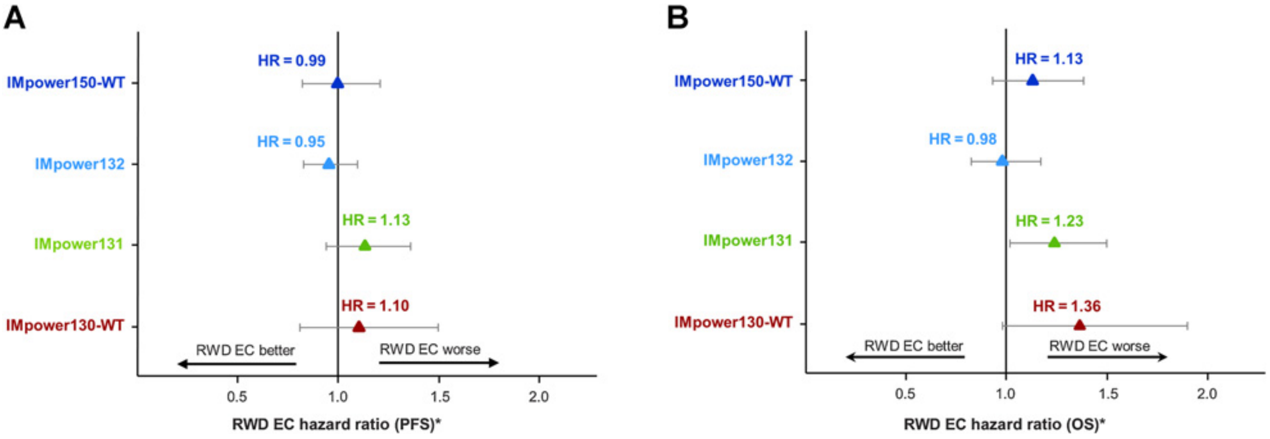

Figure 4.

Replication of PFS and OS of control arms of IMpower RCTs using RWD. HRs of PFS (**A**) and OS (**B**) of RCT control versus RWD EC arms in the IMpower150-WT, IMpower130-WT, IMpower131, and IMpower132 RCTs. EC, external control. *, Reference group is the RCT control arm.

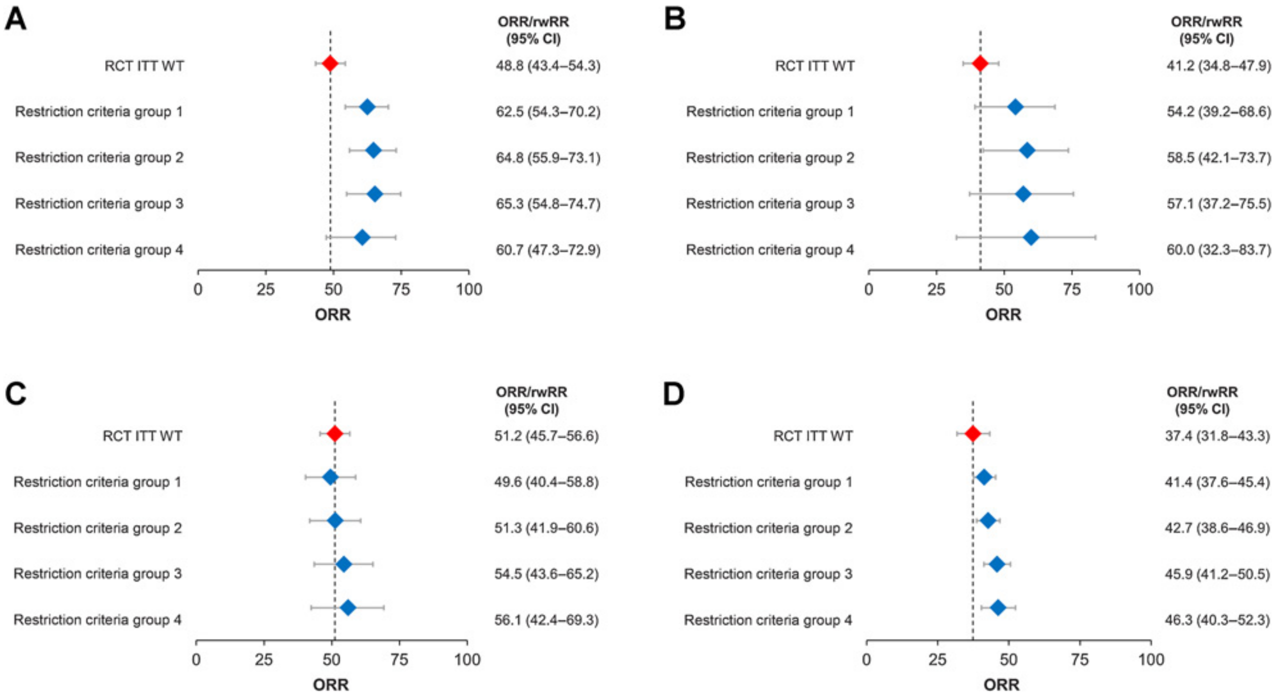

Figure 5.

RCT ORR and unadjusted rwRR according to restrictive criteria group. rwRR of RCT control versus RWD EC arms in the IMpower150-WT (**A**), IMpower130-WT (**B**), IMpower131 (**C**), and IMpower132 RCTs (**D**). EC, external control; ITT, intent to treat.

observed in some, but not all, studies. A study by Feinberg and colleagues (23) using EHR-derived information from Cardinal Health Oncology Provider Extended Network among patients with metastatic melanoma, metastatic breast cancer, or metastatic undifferentiated thyroid cancer showed that RRs categorized by physician abstraction from the EHR narrative were higher than those obtained by real-world RECIST-based measurements of lesions derived from imaging reports, with notable overestimation in the CR category. Similar findings of an overestimation of response by medical record review was also reported in a smaller study among patients with metastatic Merkel cell carcinoma (24). In other studies, rwRR provided similar estimates to ORR. In metastatic breast cancer, estimates of rwRR were similar to the ORR in the PALOMA-2 trial (22). Notably, in a recent study by Ma and colleagues (17) in which 12 different cohorts were evaluated, rwRR estimates closely mimicked trial ORR estimates. In this study, although the point estimates for rwRR were higher than ORR in IMpower150-WT, none of the comparisons between any rwRR and its corresponding trial ORR was statistically significant. Studies conducted by Ma and colleagues (17), Huang Bartlett and colleagues (22), as well as this current study all used the same abstraction methods developed by FH.

Another notable observation was that rwRR estimates were not substantively affected by population restrictions based on demographic and clinical characteristics. Furthermore, after weighting of the rwRR in the final analytic cohort to achieve balance for baseline characteristics between RWD external control and corresponding RCT control arms, points estimates did not change and all 95% CIs overlapped. Similar findings were reported in the replication of the letrozole control arm of the PALOMA-2 trial in which the rwRR did not change following IPTW (22). In another recently published study in which rwRR estimates were compared in trial-like RWD cohorts to ORR in seven different trials (12 cohorts), unweighted and weighted confirmed rwRRs were comparable (17). Collectively, the evidence suggests that typical demographic characteristics such as age, sex, ECOG PS, and stage of initial diagnosis are not important confounding variables for the treatment response endpoint. These observations are consistent with the nature of the response endpoint, which was conceptualized and developed to evaluate the impact on the tumor attributed specifically to anticancer therapeutic agents, as opposed to other clinical factors that may be more reflective of overall prognosis (25).

Despite successful replications with the rwPFS and OS endpoints, some limitations to the approach are worth noting. First, no prior data exist to validate how rwPFS is measured against RECIST-PFS within the same patient; therefore, the precise magnitude and impact of measurement error at the patient level remains unknown. Furthermore, a difference in cadence of visits and frequencies of scans between routine clinical practice and trial settings is likely to exist given the nature of these respective settings. Compared with trials that schedule follow-up scans every 6 to 8 weeks or every two cycles of therapy, clinical guidelines by the National Comprehensive Cancer Network for advanced NSCLC are described minimally as “timing of CT scans within Guidelines parameters is a clinical decision” or lacking altogether (26). In a feasibility study of patients with advanced NSCLC who had multiple radiologic assessments within the FH network, the median time between consecutive assessments was between 2 and 3 months across different treatment lines, which is less frequent than typical schedules in metastatic NSCLC trials (22, 27). The implication is, with all else being equal, a lengthier time between scans within the real-world setting is likely to produce, on average, a longer rwPFS as events are detected later. In a simulation study conducted by Adamson and colleagues, HRs for rwPFS differed from the true HRs by less than 10% in all simulated scenarios (27). We did not empirically conduct a sensitivity analysis to assess the impact of censoring at the beginning of visit gaps of various definitions (e.g., <90 days, >90 days). However, based on internal work in the OAK trial (data not shown), results for rwPFS did not substantively change when gap definitions changed from 0 days to 90 days (22, 27). In addition, for rwPFS as well as OS, statistical limitations exist. HR estimates are subject to instability early in trials when the proportion of censoring is still high or when the data are not yet fully mature (28). Finally, the generalizability of these results may be limited to only chemotherapy. As the standard of care rapidly changes to cancer immunotherapy, the ability of RWD to replicate immunotherapy control arms must be considered separately. These analyses were conducted using data sets available at a clinical cut-off date before an interim analysis, and results may have differed if more mature trial data were used.

While many potential sources of bias were minimized, differences in follow-up visit and scans frequencies, as well as immaturity of the data cut used for the RCT, could not be accounted for. Furthermore, unlike OS, PFS and ORR endpoints are inherently measured differently between RWD and trial settings. While real-world RECIST-based measurements of PFS and ORR are feasible using EHRs (23), they are still not the same as RECIST-based protocols in trials. Despite these limitations, this study also had notable strengths. First, careful consideration was taken to mimic the clinical RCT population. In the cohort selection phase, all patients meeting eligible criteria were included even if their outcomes were missing. This was particularly salient in the analysis for rwRR, in which patients with missing or indeterminate response outcomes were included as part of the denominator rather than excluded from the cohort altogether, which aligned closely with the protocols for the IMpower RCTs. In the analytic phase, careful attention was paid to defining the proper estimand and by using the standardized mortality ratio weighting methods without trimming the trial data. Finally, unlike other trial replication efforts that only had access to aggregate data (7) or reconstructed clinical trial data by digitizing Kaplan–Meier curves (19), access to patient-level data in both the RWD control arms and the RCT control arms was available, allowing for the implementation of IPTW methods and avoiding statistical vulnerabilities inherent in matching techniques that rely on aggregate data summaries (29). One of the advantages of using RWD is greater ethnic representation relative to trial participants. In fact, Flatiron data have similar sex and geographic distributions to Surveillance, Epidemiology, and End Results (SEER) as well as National Program of Cancer Registries (NPCR; ref. 12). Nonetheless, inherent limitations to the external validity of Flatiron data remain as patients are not randomly selected into the Flatiron network. Because the objective of this study was to use Flatiron data to emulate the clinical trial participants, who are highly selected, the lack of external validity does not pose a limitation.

In conclusion, this study provides support for use of RWD to provide external controls that can reasonably replicate the OS of control arms of recent first-line NSCLC RCTs. For the same RCTs, these findings also demonstrated that rwPFS abstracted from RWD can be used to replicate PFS of trial RCT arms remarkably well in first-line advanced NSCLC. Finally, an exploratory assessment of rwRR in RWD revealed a slight but nonsignificant overestimation of trial ORR, which appears to be unconfounded by baseline demographics and clinical characteristics. Extending this work to different treatments, tumor types, lines of therapy, and a thorough exploration of differences in scan frequencies in different settings will help provide greater insight into the utility of RWD in drug development.

Authors' Disclosures

N. Pal reports personal fees from Genentech during the conduct of the study, as well as personal fees from Genentech outside the submitted work. M.T. Bretscher reports employment by F. Hoffmann-La Roche (including share ownership). M.W. Lu reports employment by Genentech from 2009 to 2021, and owns shares in Roche Holding AG. G.J. Riely reports non-financial support from Genentech during the conduct of the study; G.J. Riely also reports grants from Roche, Novartis, Mirati, Takeda, and Pfizer outside the submitted work. No disclosures were reported by the other authors.

Authors' Contributions

T.G.N. Ton: Conceptualization, formal analysis, investigation, writing–original draft, writing–review and editing. **N. Pal:** Writing–original draft, writing–review and editing. **H. Trinh:** Formal analysis, writing–original draft, writing–review and editing. **S. Mahrus:** Conceptualization, investigation, writing–original draft, writing–review and editing. **M.T. Bretscher:** Conceptualization, investigation, writing–original draft, writing–review and editing. **R.J.M. Machado:** Formal analysis, writing–review and editing. **N. Sadetsky:** Conceptualization, investigation, writing–original draft, writing–review and editing. **N. Chaudhary:** Formal analysis, writing–review and editing. **M.W. Lu:** Conceptualization, investigation, writing–review and editing. **G.J. Riely:** Investigation, writing–original draft, writing–review and editing.

Acknowledgments

This manuscript was sponsored by F. Hoffmann-La Roche Ltd. and Genentech, Inc. Support for third-party writing assistance, furnished by Health Interactions, Inc., was provided by F. Hoffmann-La Roche Ltd.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received February 11, 2022; revised April 22, 2022; accepted May 3, 2022; published first May 5, 2022.

References

1. Khozin S, Blumenthal GM, Pazdur R. Real-world data for clinical evidence generation in oncology. J Natl Cancer Inst 2017;109.
2. Congress.gov. H.R.34 - 21st Century Cures Act. Available from: https://www.congress.gov/bill/114th-congress/house-bill/34.
3. US Food and Drug Administration. Framework for FDA's real-world evidence program. 2018. Available from: https://www.fda.gov/media/120060/download.
4. Arondekar B, Duh MS, Bhak RH, DerSarkissian M, Huynh L, Wang K, et al. Real-world evidence in support of oncology product registration: a systematic review of new drug application and biologics license application approvals from 2015–2020. Clin Cancer Res 2022;28:27–35.
5. Wedam S, Fashoyin-Aje L, Bloomquist E, Tang S, Sridhara R, Goldberg KB, et al. FDA approval summary: palbociclib for male patients with metastatic breast cancer. Clin Cancer Res 2020;26:1208–12.
6. Friends of Cancer Research. A blueprint for breakthrough: exploring utility of real-world evidence (RWE). Available from: https://friendsofcancerresearch.org/event/a-blueprint-for-breakthrough-exploring-utility-of-real-world-evidence-rwe/.
7. Carrigan G, Whipple S, Capra WB, Taylor MD, Brown JS, Lu M, et al. Using electronic health records to derive control arms for early phase single-arm lung cancer trials: proof-of-concept in randomized controlled trials. Clin Pharmacol Ther 2020;107:369–77.
8. Socinski MA, Jotte RM, Cappuzzo F, Orlandi F, Stroyakovskiy D, Nogami N, et al. Atezolizumab for first-line treatment of metastatic non-squamous NSCLC. N Engl J Med 2018;378:2288–301.
9. West H, McCleod M, Hussein M, Morabito A, Rittmeyer A, Conter HJ, et al. Atezolizumab in combination with carboplatin plus nab-paclitaxel chemotherapy compared with chemotherapy alone as first-line treatment for metastatic

non-squamous non–small cell lung cancer (IMpower130): a multicenter, randomized, open-label, phase III trial. Lancet Oncol 2019;20:924–37.

10. Jotte R, Cappuzzo F, Vynnychenko I, Stroyakovskiy D, Rodríguez-Abreu D, Hussein M, et al. Atezolizumab in combination with carboplatin and nab-paclitaxel in advanced squamous NSCLC (IMpower131): results from a randomized phase III trial. J Thorac Oncol 2020;15:1351–60.
11. Papadimitrakopoulou V, Cobo M, Bordoni R, Dubray-Longeras P, Szalai Z, Ursol G, et al. OA05.07 IMpower132: PFS and safety results with 1L atezolizumab + carboplatin/cisplatin + pemetrexed in stage IV non-squamous NSCLC. J Thorac Oncol 2018;13:S332–S3.
12. 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.
13. Birnbaum B, Nussbaum N, Seidle-Rathkopf K, Agrawal M, Estevez M, Estola E, et al. Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. arXiv 2020.
14. Singal G, Miller PG, Agarwala V, Li G, Kaushik G, Backenroth D, et al. Association of patient characteristics and tumor genomics with clinical outcomes among patients with non–small cell lung cancer using a clinicogenomic database. JAMA 2019;321:1391–9.
15. Curtis MD, Griffith SD, Tucker M, Taylor MD, Capra WB, Carrigan G, et al. Development and validation of a high-quality composite real-world mortality endpoint. Health Serv Res 2018;53:4460–76.
16. Zhang Q, Gossai A, Monroe S, Nussbaum NC, Parrinello CM. Validation analysis of a composite real-world mortality endpoint for patients with cancer in the United States. Health Serv Res 2021;56:1281–7.
17. Ma X, Bellomo L, Magee K, Bennette CS, Tymejczyk O, Samant M, et al. Characterization of a real-world response variable and comparison with RECIST-based response rates from clinical trials in advanced NSCLC. Adv Ther 2021;38:1843–59.
18. Brookhart MA, Wyss R, Layton JB, Stürmer T. Propensity score methods for confounding control in nonexperimental research. Circ Cardiovasc Qual Outcomes 2013;6:604–11.
19. Tan K, Bryan J, Segal B, Bellomo L, Nussbaum N, Tucker M, et al. Emulating control arms for cancer clinical trials using external cohorts created from electronic health record-derived real-world data. Clin Pharmacol Ther 2022; 111:168–78.
20. Stewart M, Norden AD, Dreyer N, Henk HJ, Abernethy AP, Chrischilles E, 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.
21. Griffith SD, Tucker M, Bowser B, Calkins G, Chang CJ, Guardino E, 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;36:2122–36.
22. Huang Bartlett C, Mardekian J, Cotter MJ, Huang X, Zhang Z, Parrinello CM, et al. Concordance of real-world versus conventional progression-free survival from a phase III trial of endocrine therapy as first-line treatment for metastatic breast cancer. PLoS One 2020;15:e0227256.
23. Feinberg BA, Zettler ME, Klink AJ, Lee CH, Gajra A, Kish JK. Comparison of solid tumor treatment response observed in clinical practice with response reported in clinical trials. JAMA Netw Open 2021;4:e2036741.
24. 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;14:2841–8.
25. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, et al. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. J Natl Cancer Inst 2000;92:205–16.
26. Ettinger DS, Wood DE, Aisner DL, Akerley W, Bauman JR, Bharat A, et al. NCCN guidelines insights: non–small cell lung cancer, version 2.2021. J Natl Compr Canc Netw 2021;19:254–66.
27. Adamson BJS, Ma X, Griffith SD, Sweeney EM, Sarkar S, Bourla AB. Differential frequency in imaging-based outcome measurement: bias in real-world oncology comparative-effectiveness studies. Pharmacoepidemiol Drug Saf 2022;31:46–54.
28. Betensky RA. Measures of follow-up in time-to-event studies: Why provide them and what should they be? Clin Trials 2015;12:403–8.
29. Signorovitch JE, Sikirica V, Erder MH, Xie J, Lu M, Hodgkins PS, et al. Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. Value Health 2012;15:940–7.