Emulating Control Arms for Cancer Clinical Trials Using External Cohorts Created From Electronic Health Record-Derived Real-World Data

Katherine Tan[1,*], Jonathan Bryan[1], Brian Segal[1], Lawrence Bellomo[1], Nate Nussbaum[1], Melisa Tucker[1], Aracelis Z. Torres[1], Carrie Bennette[1], William Capra[2], Melissa Curtis[1] and Rebecca A. Miksad[1]

Electronic health record (EHR)-derived real-world data (RWD) can be sourced to create external comparator cohorts to oncology clinical trials. This exploratory study assessed whether EHR-derived patient cohorts could emulate select clinical trial control arms across multiple tumor types. The impact of analytic decisions on emulation results was also evaluated. By digitizing Kaplan–Meier curves, we reconstructed published control arm results from 15 trials that supported drug approvals from January 1, 2016, to April 30, 2018. RWD cohorts were constructed using a nationwide EHR-derived de-identified database by aligning eligibility criteria and weighting to trial baseline characteristics. Trial data and RWD cohorts were compared using Kaplan–Meier and Cox proportional hazards regression models for progression-free survival (PFS) and overall survival (OS; individual cohorts) and multitumor random effects models of hazard ratios (HRs) for median endpoint correlations (across cohorts). *Post hoc*, the impact of specific analytic decisions on endpoints was assessed using a case study. Comparing trial data and weighted RWD cohorts, PFS results were more similar (HR range = 0.63–1.18, pooled HR = 0.84, correlation of median = 0.91) compared to OS (HR range = 0.36–1.09, pooled HR = 0.76, correlation of median = 0.85). OS HRs were more variable and trended toward worse for RWD cohorts. The *post hoc* case study had OS HR ranging from 0.67 (95% confidence interval (CI): 0.56–0.79) to 0.92 (95% CI: 0.78–1.09) depending on specific analytic decisions. EHR-derived RWD can emulate oncology clinical trial control arm results, although with variability. Visibility into clinical trial cohort characteristics may shape and refine analytic approaches.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ Electronic health record (EHR)-derived real-world data (RWD) has been evaluated as external comparator cohorts to complement clinical trials; however, prior studies focused on single tumor types and produced varying results.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ For different types of cancer, how do outcomes for patients receiving standard treatment (real-world patients) compare with patients treated with standard therapy in clinical trials (control-arm patients)?

☑ How do decisions about which real-world patients to include in the comparison with clinical-trial patients impact these results?

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ This study suggests that EHR-derived RWD can emulate clinical trial control arms across tumor types, although there may be variability in how outcomes compare. Knowing details about the clinical-trial patients may be important to identify the most appropriate real-world patients for the comparison.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ This study suggests that evaluations using EHR-derived RWD for external comparator cohorts may complement clinical trial results; however, further work is needed to define the optimal approach.

Contextualizing drug efficacy data from single-arm and small randomized clinical trials (RCTs) using robust external data sources and analytical methodologies is critical, especially in the regulatory approval setting for treatment of diseases that are rare or have high unmet medical need.[1–3] Trial-based historical external comparators or summary trial data from the literature, whereas

[1]Flatiron Health, Inc., New York, New York, USA; [2]Genentech, South San Francisco, California, USA. *Correspondence: Katherine Tan (ktan@flatiron.com)

Received April 14, 2021; accepted June 17, 2021. doi:10.1002/cpt.2351

potentially nearly contemporaneous, can be biased due to rapidly changing standards of care or underlying patient population differences.[4] As an alternative, real-world external comparator cohorts based on real-world data (RWD) have generated significant interest. However, much is unknown about optimal creation and performance of this approach.

In oncology, RWD have been used to generate evidence for improving patient outcomes, patient safety, and value in cancer care delivery.[5] Demand for increased trial and drug clinical development efficiency (i.e., time and resources) has motivated the evolution of the clinical evidence generation paradigm.[6] High-quality longitudinal electronic health record (EHR)-derived databases from which clinically relevant real-world cohorts can be constructed, may provide robust comparisons for clinical trials with nonrandomized designs.[7]

Previous studies have investigated the similarities between results obtained from traditional RCTs and well-designed studies using high-quality EHR-derived RWD.[5,8–10] However, such studies have focused on specific tumor types and may not be generalizable. Broader efforts across multiple tumor types may provide a valuable perspective about high throughput approaches and universal analytic features vs. disease-specific considerations and modifications. In addition, as differing analytic design choices have been shown to yield different results from the same data source in comparative effectiveness studies, we sought to evaluate the impact of specific analytic decisions when constructing real-world comparator cohorts.[11]

Indeed, confidence in using RWD to support clinical trials has been constrained by concerns about appropriate data and methods to be used for different contexts.[12] For example, changing standards of care over time, bias due to differences in the underlying clinical trial and RWD patient populations, and endpoint choice, may impact the recency and relevance of RWD.[13] Furthermore, inadequate (or infeasible) application of cohort eligibility criteria and insufficiently robust analytic methods could exacerbate known limitations of RWD.[9] In order to apply RWD to construct comparator cohorts in support of clinical trials, the data, design, and analysis of RWD toward emulating the target trial need to be carefully considered.[7,14]

The objective of this exploratory study was to assess the degree to which longitudinal data from a curated EHR-derived de-identified database of patients with cancer could emulate the endpoints (overall survival (OS) and progression-free survival (PFS)) observed in the control arms of selected oncology clinical trials supporting the US Food and Drug Administration (FDA)-approvals across multiple tumor types. *Post hoc*, we also evaluated the impact of specific real-world cohort construction analytic decisions on observed endpoints.

METHODS

Data sources

This study used EHR-derived data from the nationwide Flatiron Health database, a longitudinal, de-identified database containing patient-level structured data and variables curated from unstructured data via technology-enabled abstraction involving trained human curators following standardized policies and procedures.[15,16] At the time of this study, the de-identified data were derived from ~ 280 US cancer clinics (~ 800 sites of care). The majority of patients in the database originate from community oncology settings; relative community/academic proportions may vary depending on study cohort (see details of site and patient characteristics in Ma *et al.*[15]).

Patients diagnosed with advanced non-small cell lung cancer (NSCLC), metastatic breast cancer, metastatic renal cell carcinoma, advanced urothelial cancer, and multiple myeloma were included. The eligibility timeperiod for each RWD cohort varied corresponding to clinical trial cohort entry dates; the data cutoff (i.e., the date ending RWD patient observation and retrospective data collection) for all RWD cohorts was July 31, 2019 (**Table 1**). The RWD were drawn from the standard scaled data models for included tumor types; no additional trial-specific data abstraction was performed to supplement the existing "off-the-shelf" EHR-derived dataset. Institutional review board approval of the study protocol was obtained prior to study conduct, and included a waiver of informed consent.

In order to identify clinical trial control arms for potential emulation, drug approvals from January 1, 2016, to April 30, 2018, were searched to select trials with active control arms that were used to support an FDA approval.[17] This time period for drug approvals was chosen as likely to be supported by trials with enrollment periods overlapping the available EHR-derived datasets. Forty-nine RCTs that supported FDA approvals of anticancer therapies from January 1, 2016, to April 30, 2018, were selected. Drug labels and FDA announcements were reviewed to identify the trial(s) associated with each approval. We identified whether the approvals were supported by the following: at least one RCT or all single-arm studies; if randomized, whether the trial had an active or placebo (or "observation") control; and whether the FDA approval was for a cancer type available within the Flatiron Health database. We excluded trials with only placebo or "observation" controls, as it was judged infeasible to include patients receiving comparable care in real-world settings and, for "observation only patients," to assign an index date to the appropriate time point. From this refined list of 36 trials, we selected all trials in tumors that aligned with one of the existing disease-specific datasets available in the real-world database for analysis at the time of study initiation, leaving 21 trials. Six of the remaining trials were not feasible to replicate due to insufficient patients in the real-world cohorts (i.e., an initial cohort size estimate of < 25). Small cohort sizes were mostly driven by the trial control arm therapy not aligning with treatment patterns observed in the EHR-derived dataset or the use of specific biomarkers for trial eligibility not commonly tested in routine practice at that time (see the "Trial Data and RWD Cohort Construction" section below). A total of 15 trials were ultimately used for control arm replications.[18–32] A summary of the clinical trial control arms that were feasible to replicate using RWD and corresponding real-world cohort sample sizes is presented in **Table 1**.

Trial data and RWD cohort construction

Analyses were conducted according to a prespecified Statistical Analysis Plan unless otherwise noted, using a comprehensive and high-throughput approach. Trial data were reconstructed based on data presented in primary trial manuscripts by digitizing published survival curves[33] to obtain estimated patient-level endpoints and by using reported baseline characteristics as cohort-level patient covariates (henceforth named as "reconstructed trial data"). Real-world comparator cohorts were constructed from patient-level EHR-derived real-world databases and based on publicly available information or information on file in trial protocols and publications (henceforth named as "RWD cohort").[18–32,34]

Each RWD cohort was constructed in a three-step process. First, we selected patients from the existing RWD who received therapy consistent with the trial's control arm (henceforth named as "broad cohort"). Next, we aligned cohorts according to trial eligibility (inclusion/exclusion) criteria (henceforth named as "aligned cohort"). Two experienced medical oncology physician-researchers independently conducted a qualitative review of eligibility criteria to assess (i) whether deriving variables from RWD was feasible (e.g., availability in the RWD database, ability to align RWD with the trial definition), and (ii) the fitness for use in our analysis (i.e., extent of data missingness). A third senior medical

Table 1 Summary of clinical trial control arms feasible to replicate using de-identified EHR-derived RWD

Trial	Clinical condition	Control arm therapy	OS and/ or PFS endpoint(s)	Trial enrollment start[a]	Trial enrollment end[a]	Trial control arm sample size	Cohort A: Real-world patients receiving care consistent with control arm, *N*	Cohort B: Real-world cohorts after aligning with trial's eligibility criteria, *N*	Cohort C: Effective sample size of real-world cohorts after weighting, *N*
OAK	Locally advanced or metastatic NSCLC after platinum-containing CT failure	Docetaxel	OS, PFS	03/11/2014	04/29/2015	425	562	306	280.78
POPLAR	Locally advanced or metastatic NSCLC after platinum-containing CT failure	Docetaxel	OS, PFS	08/05/2013	03/31/2014	143	329	156	135.77
ALEX	Previously untreated, advanced ALK+ NSCLC	Crizotinib	OS, PFS	08/18/2014	01/20/2016	151	230	208	188.6
AURA-3	T790^{mt} advanced NSCLC with progression after 1L EGFR-TKI therapy	Pemetrexed plus either carboplatin or cisplatin	PFS	08/15/2014	09/2015[b]	140	NA[c]	NA[c]	NA[c]
KEYNOTE-021	CT-naive advanced nonsquamous NSCLC without targetable EGFR or ALK genetic aberrations	Carboplatin and pemetrexed	OS, PFS	11/25/2014	01/25/2016	63	1272	799	677.61
METEOR	Advanced or metastatic clear-cell RCC previously treated with ≥ 1 VEGFR TKIs	Everolimus	OS, PFS[d]	08/08/2013	11/24/2014	328	99	61	57.07
CABOSUN	Previously untreated metastatic RCC	Sunitinib	OS, PFS[d]	07/09/2013	04/06/2015	78	596	260	181.29
CheckMate-214	Previously untreated advanced or metastatic RCC, intermediate or poor prognostic risk	Sunitinib	OS, PFS[d]	10/01/2014	02/2016[b]	422	729	316	273.75
NCT01136733	2L treatment for metastatic RCC	Everolimus	OS, PFS[d]	03/16/2012	06/19/2013	50	49	29	26.07
POLLUX	MM after ≥ 1L of therapy	Lenalidomide and dexamethasone	OS, PFS	06/16/2014	07/14/2015	283	54	42	NA[c]

(Continued)

Table 1 (Continued)

Trial	Clinical condition	Control arm therapy	OS and/or PFS endpoint(s)	Trial enrollment start[a]	Trial enrollment end[a]	Trial control arm sample size	Cohort A: Real-world patients receiving care consistent with control arm, *N*	Cohort B: Real-world cohorts after aligning with trial's eligibility criteria, *N*	Cohort C: Effective sample size of real-world cohorts after weighting, *N*
CASTOR	MM after ≥ 1L of therapy	Bortezomib and dexamethasone	OS, PFS	09/04/2014	09/24/2015	247	NA[c]	NA[c]	NA[c]
KEYNOTE-045	Advanced UC that recurred or progressed after platinum-based CT	Investigator's choice of chemotherapy with paclitaxel, or docetaxel, or vinflunine[e]	OS, PFS	11/05/2014	11/13/2015	272	166	114	71.59
PALOMA-2	Postmenopausal women with ER+, HER2– advanced BC	Letrozole (plus placebo)	PFS	02/28/2013[b]	07/2014[b]	222	429	304	186.49
MONALEESA-2	Postmenopausal women with HR+, HER2- recurrent or metastatic BC	Letrozole (plus placebo)	PFS	01/24/2014	03/24/2015	334	542	334	257.12
MONARCH-3	Postmenopausal women with HR+, HER2- advanced BC	Anastrozole or letrozole (plus placebo)	PFS	11/18/2014	11/11/2015	165	1,324	926	561.25

Sample sizes for the corresponding real-world cohorts: (A) all patients receiving control arm therapy in real-world settings, (B) after aligning the real-world cohort with the trial's eligibility criteria, and (C) after weighting the aligned real-world cohort to balance differences in key prognostic factors between the control arm and real-world patients.

1L, first line; 2L, second line; ALK, anaplastic lymphoma kinase; BC, breast cancer; CT, chemotherapy; EGFR, epidermal growth factor receptor; EHR, electronic health record; ER, estrogen receptor; HER, human epidermal growth factor receptor; HR, hormone receptor; MM, multiple myeloma; NA, not applicable; NSCLC, non-small cell lung cancer; OS, overall survival; PFS, progression-free survival; RCC, renal cell carcinoma; RWD, real-world data; TKI, tyrosine kinase inhibitor; UC, urothelial cancer; VEGFR, vascular endothelial growth factor receptor.

[a] As publicly reported. [b] Exact date not reported (assumed end of month). [c] Cohorts with N < 25 were not included due to small sample sizes. [d] Endpoint not (yet) available from Flatiron Health data at scale.

[e] Vinflunine not marketed in the United States and therefore not relevant when selecting US-based real-world cohorts.

oncologist was pre-identified as an adjudicator in case of disagreement that could not be resolved. Not all clinical trial criteria were relevant to or available as part of routine care; eligibility criteria relied on data (structured and unstructured) that were available in the existing EHR-derived RWD at the time of analysis. Eligibility criteria were documented as being implemented (yes/no) in the RWD cohort construction process and further categorized according to implementations status (**Table S1**). Finally, we reduced any observed imbalances in baseline characteristics between trial and real-world cohorts through a covariate balancing method analogous to propensity score weighting (henceforth named as "weighted cohort"). For each trial, baseline characteristics for weighting were drawn from summary-level statistics in the primary trial publications: age, gender, race, smoking history, and histology (e.g., squamous vs. non-squamous NSCLC) as applicable, as well as tumor-specific variables (e.g., prior nephrectomy for patients with metastatic renal cell carcinoma or International Staging System stage for multiple myeloma) if deemed clinically relevant (**Table S2**). Then, inverse-odds weights were estimated for real-world patients using the generalized method of moments estimator on trial reported moments (i.e., means or proportions of variables) such that the weighted RWD cohorts and reconstructed trial data achieved cohort-level balance for each included baseline characteristics marginally.[35,36] The main analysis of the aligned and weighted RWD cohorts included all patients with missing Eastern Cooperative Oncology Group (ECOG) performance status or laboratory tests because real-world capture of this information is often incomplete, as well as patients who received the control arm therapy of interest prior to the start of trial enrollment date (i.e., "historical control therapies") assuming little change in standard of care therapies; sensitivity analyses investigated the impact of excluding these patient groups.

Endpoints

As described above, each reconstructed trial dataset was associated with three corresponding RWD cohorts ("broad," "aligned," and "weighted"). Comparison for each trial-RWD pair (3 RWD cohorts for each of the 15 trials) was based on the following endpoints: PFS and OS as reported in the primary publications for each trial, and real-world PFS (to the extent available in the existing RWD) and real-world OS (rwOS) for the RWD cohorts. The real-world PFS and rwOS endpoints were derived from real-world progression and mortality variables, described previously in the literature.[37,38] PFS and OS were analyzed for all RWD-trial pair comparisons regardless of whether the clinical trial primary endpoint was OS, PFS, or another endpoint.

Analyses

Comparing endpoints in individual trials. To ensure that follow-up time between the trial-RWD pairs were comparable, we censored the RWD cohorts so that the maximum potential follow-up duration for real-world endpoint evaluation was equal to the maximum trial endpoint follow-up duration as reported in the primary trial publication. For trials that censored PFS at the time of the start of a subsequent anticancer therapy, we applied analogous endpoint censoring procedures to the corresponding real-world cohorts, as feasible with RWD. For each trial-RWD pair, we performed visual inspection on Kaplan–Meier curves and computed median times-to-event separately for the reconstructed trial data and the RWD cohort. Then, we computed relative time-to-event estimates comparing reconstructed trial data vs. RWD cohort using Cox proportional hazards regression models separately on the broad, aligned, and weighted RWD cohorts, as described above. For the Cox models, hazard ratios (HRs) = 1 indicated comparable endpoints, HRs < 1 indicated hazard of outcomes (death or progression) was lower in the reconstructed trial data compared to RWD cohorts, and HRs > 1 indicated that the hazard of outcomes (death or progression) was higher in the reconstructed trial data compared to RWD cohorts.

Comparing endpoints across all trials. Endpoint comparisons were done for the trial-RWD pairs with weighted RWD cohorts only, separately for the OS and PFS endpoints. Specifically, we conducted meta-analyses across all trials to obtain pooled HR estimates using a multitumor random effects model, and assessed heterogeneity using the I2 statistic.[39] In addition, *post hoc* correlation plots of median time-to-event endpoints (OS and PFS) for trial-RWD pairs were created.

Post hoc assessment on impact of specific analytic decisions

Post hoc, we assessed the impact of specific analytic decisions on endpoints and overall results using a case study. After initiation of study conduct, a patient-level analysis of multiple advanced NSCLC trials by Carrigan *et al*. reported detailed information on the replication approach for the OAK trial (an open-label, randomized phase III study of atezolizumab vs. docetaxel in adult patients with squamous or nonsquamous advanced NSCLC).[8] Therefore, we modified the weighted RWD cohort construction for the OAK trial in our study by retrofitting analytic decisions that differed from those described in Carrigan *et al*.: (i) updated to the same RWD cutoff date, (ii) applied eligibility criteria using the same subset of laboratory tests, (iii) included RWD patients who received therapy after trial enrollment ended, and (iv) included time from initial diagnosis to therapy start (identified in the Carrigan *et al*. paper as an important factor through analysis of the trial patient-level data) as a weighing variable. Finally, we evaluated the OS HR of the resulting trial-RWD pairs.

RESULTS

Construction of real-world comparator cohorts

Details on the feasibility of implementing eligibility criteria for each included trial are presented in **Table S1**. A summary of clinical trials not included in this study based on initial feasibility assessment is provided in **Table S3**. We observed a series of common eligibility criteria that could not be implemented across trials, such as process-related activities (e.g., providing tumor tissue or signing consent forms), documented assessment of prior adverse events and their resolution before index/study treatment, and subjective criteria regarding patients' expected prognoses or adherence to study-specific measures. Note that the proportion of trial criteria that could be implemented was variable, depending on factors such as disease and therapy types as well as trial-specific considerations. The majority of these types of criteria were determined as either not relevant for or not expected to be documented as a part of routine standard of care by authors with oncology clinical trial expertise; therefore, they were not included in the standard data model for this analysis.

Endpoint comparisons

An example of Kaplan–Meier curves for the reconstructed trial data and RWD cohort is shown for the OAK trial (**Figure 1**). OS and PFS HR comparisons between the trial-RWD pairs are presented in **Table 2**. Based on the weighted RWD cohorts, endpoints between the trial-RWD pairs appeared to be more similar for PFS compared to OS (HR range: OS = 0.36–1.09, PFS = 0.63–1.18) and generally variable. Comparing the selected trials by disease subtypes, endpoints appear to be most similar for patients with a urothelial cancer (HR range: OS = 0.79–1.09, PFS = 0.73–1.15) and most dissimilar for patients with advanced NSCLC, especially for OS (HR range: OS = 0.36–0.68, PFS = 0.65–0.80). Endpoint dissimilarity was most pronounced for the trial-RWD pair of KEYNOTE-021 (an open-label, phase II study of

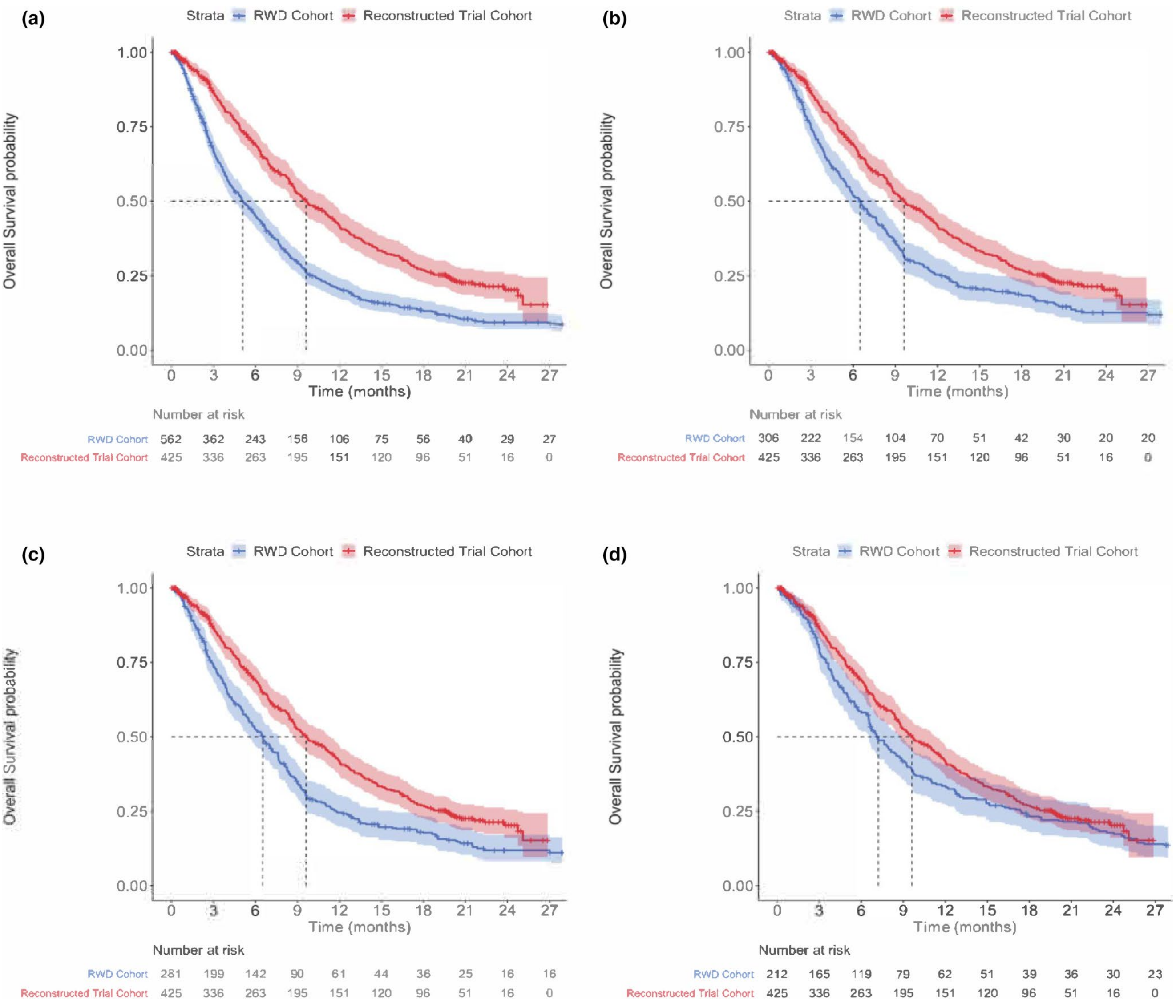

Figure 1 Comparison of overall survival curves from control arm from OAK (re-constructed) vs. real-world patients. *Notes:* (**a**) Real-world patients receiving control arm therapy; (**b**) after aligning with the trial's eligibility criteria; (**c**) after weighting the aligned real-world cohort to balance any remaining differences in key prognostic factors; (**d**) after reproducing analytic decisions from a publication using the same data source: Carrigan *et al*.[8]RWD, real-world data. [Colour figure can be viewed at wileyonlinelibrary.com]

immunotherapy for nonsquamous NSCLC without mutations that allowed cross-over at progression; HRs of OS = 0.36, PFS = 0.65).

The correlation between trial-RWD pairs for median OS and median PFS was 0.85 and 0.91, respectively (**Figure 2**). The OS correlation excluded estimates from the ALEX and KEYNOTE-021 trial-RWD pairs in advanced NSCLC trials because the median OS was not reached in the reconstructed trial data (although it was reached in the corresponding RWD cohorts). Based on meta-analytic multitumor random effects models on all available trials, the combined HR was 0.76 for OS (I2 = 63%) and 0.84 for PFS (I2 = 62%), suggesting overall poorer outcomes in the RWD cohorts compared with the reconstructed trial data and overall substantial heterogeneity across trials. Results from sensitivity analyses of constructing aligned and weighted RWD cohorts (i.e., excluding patients with missing ECOG performance status, missing laboratory tests, and receiving historical control therapies) were largely consistent with the main analyses (**Figures S1–S5**).

Post hoc assessment on impact of specific analytic decisions

The *post hoc* assessment on the impact of specific analytic decisions on replication results was based on the OAK study. The original methodology deployed for all trial-RWD pairs had an OS HR of 0.67 (95% confidence interval: 0.56–0.79) comparing reconstructed trial data to the weighted RWD cohort (**Figure 1c**). When weighted RWD cohorts were modified by retrofitting analytic decisions similar to those described in Carrigan *et al*., the resulting OS more closely aligned with that of the OAK clinical trial (OS HR = 0.92; 95% confidence interval: 0.78–1.09; **Figure 1d**); stepwise application of specific analytic decisions resulted in highly variable incremental changes in HRs (**Figure 3**). Updating to the

Table 2 Comparison of trial vs. real-world cohort outcomes

Trial	(A) Real-world patients receiving control arm therapy		(B) Aligned with trial eligibility criteria		(C)[a] Aligned and weighted	
	PFS	OS	PFS	OS	PFS	OS
	HR [95% LCL, UCL]		HR [95% LCL, UCL]		HR [95% LCL, UCL]	
OAK	0.75 [0.65, 0.87]	0.57 [0.49, 0.66]	0.82 [0.69, 0.97]	0.68 [0.58, 0.81]	0.80 [0.67. 0.95]	0.67 [0.56, 0.79] 0.92 [0.78, 1.09][b]
POPLAR	0.70 [0.57, 0.86]	0.56 [0.44, 0.70]	0.75 [0.59, 0.96]	0.68 [0.52, 0.89]	0.73 [0.57, 0.93]	0.66 [0.50, 0.87]
ALEX	0.79 [0.62, 1.01]	0.63 [0.44, 0.91]	0.79 [0.62, 1.02]	0.67 [0.46, 0.97]	0.80 [0.62, 1.02]	0.68 [0.47, 1.00]
AURA-3	N/A (*n* < 25)	N/A (*n* < 25)	N/A (*n* < 25)	N/A (*n* < 25)	N/A (*n* < 25)	N/A (*n* < 25)
KEYNOTE-021	0.60 [0.43, 0.85]	0.29 [0.17, 0.49]	0.63 [0.45, 0.89]	0.33 [0.20, 0.57]	0.63 [0.45, 0.89]	0.36 [0.21, 0.61]
METEOR	1.02 [0.78, 1.33]	0.89 [0.66, 1.20]	1.03 [0.74, 1.42]	0.93 [0.65, 1.34]	1.05 [0.75, 1.46]	1.00 [0.69, 1.45]
CABOSUN	1.27 [0.97, 1.66]	1.05 [0.76, 1.46]	1.19 [0.89, 1.58]	1.11 [0.78, 1.58]	1.15 [0.86, 1.55]	1.09 [0.76, 1.58]
CheckMate-214	0.79 [0.68, 0.93]	0.83 [0.70, 0.99]	0.68 [0.56, 0.82]	0.81 [0.65, 0.99]	0.65 [0.54, 0.79]	0.80 [0.65, 0.99]
NCT01136733	0.84 [0.54, 1.31]	0.72 [0.41, 1.24]	0.84 [0.51, 1.4])	0.85 [0.45, 1.61]	0.73 [0.43, 1.23]	0.79 [0.41, 1.52]
POLLUX	0.72 [0.45, 1.15]	0.78 [0.41, 1.50]	0.59 [0.36, 0.99]	0.83 [0.40, 1.71]	N/A (*n* < 25)	N/A (*n* < 25)
CASTOR	N/A (*n* < 25)	N/A (*n* < 25)	N/A (*n* < 25)	N/A (*n* < 25)	N/A (*n* < 25)	N/A (*n* < 25)
KEYNOTE-045	N/A	0.74 [0.59, 0.93]	N/A	0.81 [0.63, 1.05]	N/A	1.06 [0.77, 1.45]
PALOMA-2	1.01 [0.89, 1.36]	N/A	1.21 [0.96, 1.52]	N/A	1.18 [0.91, 1.55]	N/A
MONALEESA-2	1.1 [0.89, 1.35]	N/A	1.03 [0.82, 1.30]	N/A	0.94 [0.74, 1.21]	N/A
MONARCH-3	1.07 [0.86, 1.34]	N/A	1.12 [0.89, 1.41]	N/A	1.00 [0.79, 1.27]	N/A

HRs compare time-to-event endpoints of trial control versus real-world cohort patients.

HR = 1 indicates comparable endpoints, HR < 1 indicates real-world endpoints observed to be worse than trial endpoints, HR > 1 indicates real-world endpoints observed to be better than trial endpoints. Estimates are not shown when the effective sample size was < 25 patients.

HR, hazard ratio; LCL, lower confidence limit; N/A, not applicable; OS, overall survival; PFS, progression-free survival; UCL, upper confidence limit.

[a] Limitations due to lack of trial patient-level data access prevented replication of all trials using the analytic decisions as was done for the OAK trial. [b] OS HR of 0.92 for the OAK replication was computed under a different set of analytic decisions, such as including real-world patients from a different time window and an expanded set of baseline covariates; see "*Post hoc* assessment of specific analytic decisions" in Methods and Results sections for more details.

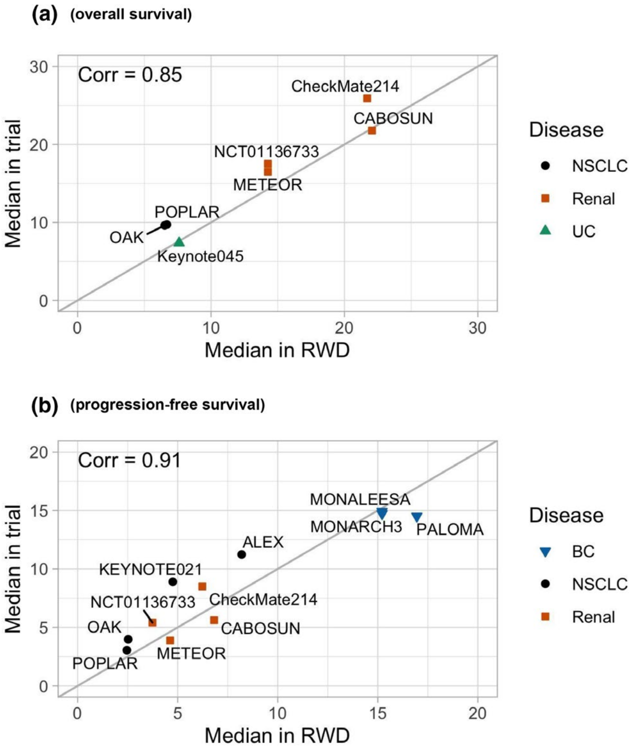

Figure 2 Correlation plot of median time-to-events comparing patients in the original trial's control arm vs. a weighted real-world cohort for (a) Overall Survival (OS) and (b) Progression Free Survival (PFS). Note: ALEX and KEYNOTE-021 (both advanced NSCLC) were excluded from the OS plot as median OS was not reached in the trials' control arms. BC, breast cancer; UC, urothelial cancer; Corr, correlation; NSCLC, non-small cell lung cancer; OS, overall survival; PFS, progression-free survival; RWD, real-world data. [Colour figure can be viewed at wileyonlinelibrary.com]

same data cutoff date for the RWD cohort (HR change = 0.03), applying the same eligibility criteria (HR change = 0.05), including RWD patients who received therapy after trial enrollment ended (HR change = 0.06), and including time from initial diagnosis to therapy start as a weighing variable (HR change = 0.11).

DISCUSSION

In this study, we used a three-step process to retrospectively construct RWD cohorts using "off-the-shelf" datasets aligned to published clinical trials that supported FDA approvals of anticancer therapies across multiple tumor types. We showed that endpoints of contemporaneous RWD cohorts are directionally similar to those of clinical trial control arms, as evidenced by the correlations of median OS and PFS. Our findings suggest RWD cohorts' endpoints appear to be variable and trend toward worse outcomes in the RWD cohorts. The *post hoc* case study investigated the impact of applying different analytic decisions and suggested that the comparability of clinical trial control arms and RWD cohorts may be impacted by variability in data source, choice and availability of relevant prognostic factors for inclusion in the statistical analysis, and RWD cohort construction analytic decisions.

Our evaluation of EHR-derived real-world cohorts as comparators for clinical trial control arms was conducted across multiple tumor types and analyzed in a meta-analytic fashion, adding to the existing body of evidence in oncology that is otherwise limited to studies of single tumor types (glioblastoma,[40] ALK+ advanced NSCLC,[10] metastatic breast cancer,[41] and gastric cancer[42]). Although other efforts have examined the feasibility of claims-derived RWD to replicate clinical trial results,[9] and explored using historical clinical trial patient-level data to develop synthetic control arms and to replicate RCT endpoints,[43] EHR-derived RWD is unique in its combination of depth, longitudinality, and contemporaneousness. Our results corroborate findings from existing literature that suggest using RWD for external controls may produce similar endpoints as observed in clinical trial control arms and also show that comparability between trial and RWD controls may be context-dependent. For example, the nuances of data capture of individual variables (which may differ among variables, sources, and databases that use different approaches to derive information

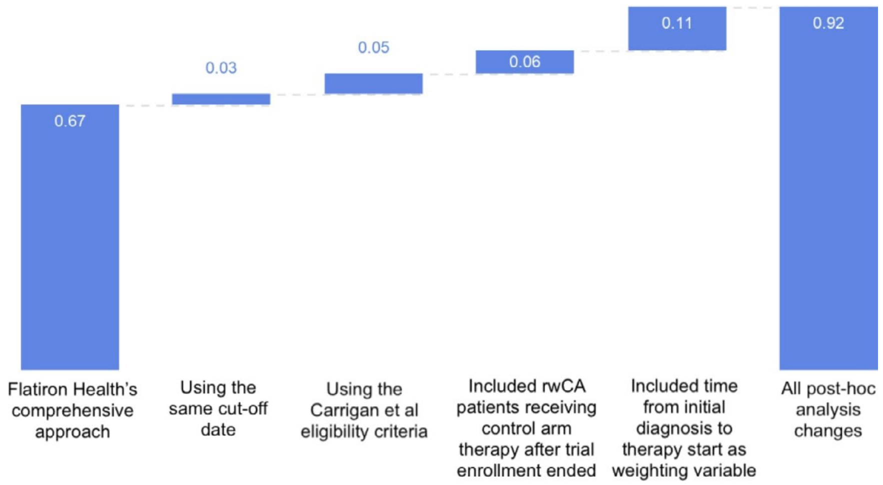

Figure 3 Waterfall plot of increase in OS HR, comparing trial to real-world control cohorts, when incremental analytic decisions were applied to the real-world cohort. *Notes:* Incremental changes in analytic decisions were based on the description in: Carrigan *et al*.[8] HR, hazard ratio; OS, overall survival; rwCA, real-world control arm. [Colour figure can be viewed at wileyonlinelibrary.com]

from the same source) may dictate specific approaches in applying clinical trial eligibility criteria to RWD cohorts. Therefore, the transparent accounting and documentation of the clinical rationale behind these decisions is critical.

As demonstrated in our line-by-line assessment of individual trial eligibility criteria, the primary barriers to implementation of specific eligibility criteria were: (i) data elements that were not available in the specific EHR-derived RWD model available at the time of the study (e.g., baseline comorbidities and nuanced disease characteristics); (ii) data elements that have limited relevance or are not routinely captured in the clinical setting when using RWD in general (e.g., trial logistics and trial-specific assessments, such as sufficient life expectancy); (iii) differing testing patterns for specific tumor mutations in the real-world setting than were required as part of screening for a clinical trial; and (iv) knowledge of trials' full eligibility criteria (as often only the major inclusion/exclusion criteria are stated in publication). The first barrier may be overcome with customized abstraction and curation of tailored data models, whereas the other barriers may not have a significant impact in some settings or are known limitations of RWD. Despite these hurdles, our approach produced high correlations of OS and PFS for trials in this study; the challenge is prospectively identifying the appropriate context for when to use this approach. Development of trial-specific RWD data models is likely an important step toward more universal application of real-world external comparators, especially for regulatory settings.

In the *post hoc* case study, we found that the degree of similarity between trial and RWD cohorts can be sensitive to different analytic cohort construction decisions resulting from varying assumptions and potential biases. For example, the case study illustrated that RWD cohort endpoints in retrospective comparison cohort construction may be impacted by (i) lack of visibility into full and relevant trial data, which may result in omission of important prognostic factors (in this case, variables related to disease aggressiveness), and (ii) inclusion of "futuristic" RWD controls, (i.e., real-world patients who started trial control therapy in the time window after trial enrollment ended). These "post enrollment patients" may have had different subsequent treatment options compared with those who started therapy during or before trial enrollment. The poorer outcomes of RWD compared with RCT cohorts under certain analytic decisions could result in treatment efficacy overestimation when replacing RCT with RWD controls. Thus, successful replication of clinical trial control arms using RWD cohorts requires selection of a comprehensive set of context-specific prognostic factors. Although we desired to explore the impact of the analytic decisions of Carrigan *et al*. applied to the other trials in our study, this was not feasible because specific variables were not publicly available. Although the reported approach in Carrigan *et al*. resulted in an OS endpoint treatment effect that closely matched that of corresponding clinical trials, their work was limited to NSCLC trials, although similar approaches could be extended to other disease contexts. Other potential methodologies include using the observed RWD to construct plausible Directed Acyclic Graphs to describe causal relationships (or lack thereof) among potential prognostic factors in relation to the endpoints of interest.[44,45] Ultimately, the range of observed differences between trial and RWD cohorts, through retrospective evaluations, such as this and other reports, may inform confidence levels for new RWD methodology and data sources used to support clinical trials. For example, these results may help prespecify reasonable bounds of the true differences in treatment effects in a threshold-crossing framework analysis[46] or plausible outcome distributions for imputation in a tipping point analysis.

This study is subject to several limitations. First, there are limits to the comparability of the EHR-derived clinical data in our study to the trial cohorts due to missing or unmeasured data, a common RWD limitation (e.g., events not documented or occurring outside of the source network).[9] Such missing or incomplete data, if clinically prognostic for the endpoints of interest, could have impacted results and interpretation due to unmeasured confounding. Related, the EHR-derived data were drawn from nationwide, predominantly community, cancer clinics in the United States;

Table 3 Summary of key study observations

Key observation
1. Factors such as variability in data source, alignment of RCT and RWD eligibility criteria and study design, choice of available prognostic factors for inclusion in the statistical analysis, and varying analytic assumptions can impact the comparability of clinical trial control arms and RWD cohorts.
2. Real-world external comparator cohorts may produce similar outcomes as observed in clinical trial control arms, however, comparability may be context-dependent.
3. Transparent accounting and documentation of the clinical rationale behind decisions to apply or not apply certain clinical trial eligibility criteria to real-world external comparators is essential.
4. Certain barriers to implementation of specific clinical trial eligibility criteria when constructing real-world external comparators may be overcome with customized abstraction and curation of tailored data models, while others remain as known limitations of RWD. Of note, some eligibility criteria may not be meaningful outside of a clinical trial setting (e.g., ability to sign consent).
5. Successful replication of clinical trial control arms using RWD requires careful selection of a comprehensive set of disease- and trial-specific prognostic factors.
6. Access to clinical trial patient-level data is important to inform appropriate RWD study design and cohort construction.
7. Prospective applications of real-world external comparators for ongoing or future trials performed in collaboration with the study sponsor should take advantage of the opportunity to use patient-level data and proactively incorporate analytic considerations such as comprehensive prognostic variable data capture.

RCT, randomized clinical trials; RWD, real world data.

representativeness of other settings may differ, for example, from clinical trials that recruit from both US and non-US centers where care and insurance coverage may impact variability in clinical outcomes. Second, we were not able to account for post-baseline characteristics in this exploratory analysis due to lack of patient-level trial data. For example, subsequent treatment options available to real-world vs. trial-control arm patients may differ (particularly if crossover to experimental treatment is allowed in the trial), which may impact longer-term endpoints such as OS. Third, in this study, ascertainment of disease progression from RWD relied on a clinician-anchored abstraction approach (previously described[35]) that differs from the Response Evaluation Criteria in Solid Tumors (RECIST)-based measurements of disease progression commonly used in clinical trials.[47] Potential differences in imaging assessment cadence could have resulted in overestimation of PFS in the real-world cohort. Fourth, this study relied on trials' primary publications for balancing on cohort-level baseline covariates and reconstruction of outcomes data. Although our summary-level inverse odds weighting approach achieved covariate balance marginally for each included variable, modeling on the aggregate level cannot capture nonlinear or additive covariate effects due to the ecological inference fallacy.[48] Furthermore, it is possible that rwOS could be underestimated relative to trial OS, as some primary publications were based on PFS (and thus trial OS may be immature), or overestimated due to missing data in real-world mortality (although we expect this effect to be minor due to the high accuracy of the real-world mortality variable used).[37] Finally, although a wide range of cancer tumor types were included in this study, generalizability beyond these treatment settings and outside of the oncology setting may be limited.

In conclusion, EHR-derived RWD is a potential data source for emulating clinical trial control arm outcomes. Our findings, which are summarized in **Table 3**, suggest that visibility into clinical trial patient-level data is a key step to allow for appropriate RWD cohort construction analytic choices. Although our retrospective trial replications aimed to account for many of these factors in alignment and weighting, we were largely limited by only having access to metrics that were publicly available. Prospective applications of RWD as an external control arm for ongoing or future trials performed in collaboration with the study sponsor should take advantage of the opportunity to evaluate patient-level factors in the trial cohort and to proactively incorporate analytic considerations, such as comprehensive prognostic variable data capture. Ideally, these steps can be accomplished in the study design stage through close collaborations with experts and stakeholders, including regulators when relevant. Lessons from this retrospective study may help refine future work that evaluates retrospectively and prospectively constructed EHR-derived RWD cohorts as external comparators for single-arm studies.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

ACKNOWLEDGMENTS

The authors would like to thank Cody Patton and Hannah Gilham of Flatiron Health, Inc. for medical writing and publication management support. They would also like to thank Ariel Bourla, Meghna Samant, and Shrujal Baxi of Flatiron Health, Inc. for their contributions to this study.

FUNDING

This study was sponsored by Flatiron Health Inc., which is an independent subsidiary of the Roche group. Writing/editing support was provided by Julia-Saiz Shimosato, Cody Patton, and Hannah Gilham from Flatiron Health, Inc.

CONFLICT OF INTEREST

K.T., J.B., B.S., L.B., N.N., M.T., A.Z.T., C.B., M.C., and R.A.M. report employment in Flatiron Health, Inc., an independent subsidiary of the Roche group, and stock ownership in Roche. B.S., L.B., N.N., M.T., A.Z.T., M.C., and R.A.M. report equity ownership in Flatiron Health, Inc. W.C. reports employment in Genentech, a member of the Roche Group, and stock ownership in Roche.

AUTHOR CONTRIBUTIONS

K.T., J.B., B.S., L.B., N.N., M.T., A.Z.T., C.B., W.C., M.C., and R.A.M. wrote the manuscript. K.T., J.B., B.S., L.B., N.N., M.T., A.Z.T., C.B., W.C., M.C., and R.A.M. designed the research. K.T., J.B., C.B., and B.S. performed the research. K.T., J.B., C.B., and B.S. analyzed the data.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study have been originated by Flatiron Health, Inc. These de-identified 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.

1. Joffe, E., Iasonos, A. & Younes, A. Clinical trials in the genomic era. *J. Clin. Oncol.* **35**, 1011 (2017).
2. Ventz, S., Alexander, B.M., Parmigiani, G., Gelber, R.D. & Trippa, L. Designing clinical trials that accept new arms: an example in metastatic breast cancer. *J. Clin. Oncol.* **35**, 3160–3168 (2017).
3. Redig, A.J. & Jänne, P.A. Basket trials and the evolution of clinical trial design in an era of genomic medicine. *J. Clin. Oncol.* **33**, 975–977 (2015).
4. Viele, K. *et al*. Use of historical control data for assessing treatment effects in clinical trials. *Pharmaceut. Stat.* **13**, 41–54 (2014).
5. Khozin, S., Blumenthal, G.M. & Pazdur, R. Real-world data for clinical evidence generation in oncology. *J. Natl. Cancer Inst.* **109**, 10 (2017).
6. Hudson, K.L. & Collins, F.S. The 21st century cures act—a view from the NIH. *N. Engl. J. Med.* **376**, 111–113 (2017).
7. US Food and Drug Administration. Framework for FDA'S real-world evidence program <https://www.fda.gov/media/120060/download> Updated 2018. Accessed February 11, 2021.
8. Carrigan, G. *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.* **107**, 369–377 (2020).
9. Fralick, M., Kesselheim, A.S., Avorn, J. & Schneeweiss, S. Use of health care databases to support supplemental indications of approved medications. *JAMA Intern. Med.* **178**, 55–63 (2018).
10. Davies, J. *et al*. Comparative effectiveness from a single-arm trial and real-world data: Alectinib versus ceritinib. *J. Comparative Effectiveness Res.* **7**, 855–865 (2018).
11. Madigan, D., Ryan, P.B. & Schuemie, M. Does design matter? Systematic evaluation of the impact of analytical choices on

effect estimates in observational studies. *Ther. Adv. Drug Saf.* **4**, 53–62 (2013).

12. Collins, R., Bowman, L., Landray, M. & Peto, R. The magic of randomization versus the myth of real-world evidence. *N. Engl. J. Med.* **382**, 674–678 (2020).
13. Raphael, M.J., Gyawali, B. & Booth, C.M. Real-world evidence and regulatory drug approval. *Nat. Rev. Clin. Oncol.* **17**, 271–272 (2020).
14. Hernán, M.A. & Robins, J.M. Using big data to emulate a target trial when a randomized trial is not available. *Am. J. Epidemiol.* **183**, 758–764 (2016).
15. Ma, X., Long, L., Moon, S., Adamson, B.J.S. & Baxi, S.S. Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron health, SEER, and NPCR. medRxiv. 2020. http://medrxiv.org/content/early/2020/05/30/2020.03.16.20037143.abstract. [e-pub ahead of print].
16. Birnbaum, B. *et al*. Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. Computer Science. https://arxiv.org/abs/2001.09765. [e-pub ahead of print].
17. US Food and Drug Administration. Drugs@FDA: FDA-approved drugs <https://www.accessdata.fda.gov/scripts/cder/daf/>. Accessed February 11, 2021.
18. Fehrenbacher, L. *et al*. Atezolizumab versus docetaxel for patients with previously treated non-small-cell lung cancer (POPLAR): a multicentre, open-label, phase 2 randomised controlled trial. *Lancet* **387**, 1837–1846 (2016).
19. Bellmunt, J. *et al*. Pembrolizumab as second-line therapy for advanced urothelial carcinoma. *N. Engl. J. Med.* **376**, 1015–1026 (2017).
20. Choueiri, T.K. *et al*. Cabozantinib versus everolimus in advanced renal cell carcinoma (METEOR): final results from a randomised, open-label, phase 3 trial. *Lancet Oncol.* **17**, 917–927 (2016).
21. Choueiri, T.K. *et al*. Cabozantinib versus sunitinib as initial targeted therapy for patients with metastatic renal cell carcinoma of poor or intermediate risk: the alliance A031203 CABOSUN Trial. *J. Clin. Oncol.* **35**, 591–597 (2017).
22. Dimopoulos, M.A. *et al*. Daratumumab, lenalidomide, and dexamethasone for multiple myeloma. *N. Engl. J. Med.* **375**, 1319–1331 (2016).
23. Finn, R.S. *et al*. Palbociclib and letrozole in advanced breast cancer. *N. Engl. J. Med.* **375**, 1925–1936 (2016).
24. Goetz, M.P. *et al*. MONARCH 3: abemaciclib as initial therapy for advanced breast Cancer. *J. Clin. Oncol.* **35**, 3638–3646 (2017).
25. Langer, C.J. *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.* **17**, 1497–1508 (2016).
26. Motzer, R.J. *et al*. Lenvatinib, everolimus, and the combination in patients with metastatic renal cell carcinoma: a randomised, phase 2, open-label, multicentre trial. *Lancet Oncol.* **16**, 1473–1482 (2015).
27. Motzer, R.J. *et al*. Nivolumab plus ipilimumab versus sunitinib in advanced renal-cell carcinoma. *N. Engl. J. Med.* **378**, 1277–1290 (2018).
28. Palumbo, A. *et al*. Daratumumab, bortezomib, and dexamethasone for multiple myeloma. *N. Engl. J. Med.* **375**, 754–766 (2016).
29. Peters, S. *et al*. Alectinib versus crizotinib in untreated ALK-positive non-small-cell lung cancer. *N. Engl. J. Med.* **377**, 829–838 (2017).
30. Rittmeyer, A. *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* **389**, 255–265 (2017).
31. Verma, S. *et al*. Health-related quality of life of postmenopausal women with hormone receptor-positive, human epidermal growth factor receptor 2-negative advanced breast cancer treated with ribociclib + letrozole: results from MONALEESA-2. *Breast Cancer Res. Treat.* **170**, 535–545 (2018).
32. Wu, Y.L. *et al*. CNS efficacy of osimertinib in patients with T790M-positive advanced non-small-cell lung cancer: data from a randomized Phase III Trial (AURA3). *J. Clin. Oncol.* **36**, 2702–2709 (2018).
33. Guyot, P., Ades, A.E., Ouwens, M.J. & Welton, N.J. Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan–Meier survival curves. *BMC Med. Res. Methodol.* **12**, 9 (2012).
34. Data on file F. Hoffmann-La Roche Ltd. A phase II, open-label multicenter, randomized study to investigate the efficacy and safety of atezolizumab (anti–PD-L1 antibody) compared with docetaxel in patients with non-small cell lung cancer after platinum failure. Clinical study protocol GO28753; December 2016. (Registered with EudraCT ID 2013-001142-34).
35. Westreich, D., Edwards, J.K., Lesko, C.R., Stuart, E. & Cole, S.R. Transportability of trial results using inverse odds of sampling weights. *Am. J. Epidemiol.* **186**, 1010–1014 (2017).
36. Segal, B.D. & Bennette, C.S. Re: "Transportability of trial results using inverse odds of sampling weights". *Am. J. Epidemiol.* **187**, 2716–2717 (2018).
37. Curtis, M.D. *et al*. Development and validation of a high-quality composite real-world mortality endpoint. *Health Serv. Res.* **53**, 4460–4476 (2018).
38. Griffith, S.D. *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.* **36**, 2122–2136 (2019).
39. Higgins, J.P. & Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21**, 1539–1558 (2002).
40. Ventz, S., Lai, A., Cloughesy, T.F., Wen, P.Y., Trippa, L. & Alexander, B.M. Design and evaluation of an external control arm using prior clinical trials and real-world data. *Clin. Cancer Res.* **25**, 4993–5001 (2019).
41. Huang Bartlett, C. *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* **15**, e0227256 (2020).
42. Chau, I. *et al*. Developing real-world comparators for clinical trials in chemotherapy-refractory patients with gastric cancer or gastroesophageal junction cancer. *Gastric Cancer* **23**, 133–141 (2020).
43. Friends of Cancer Research. Exploring whether a synthetic control arm can be derived from historical clinical trials that match baseline characteristics and overall survival outcome of a randomized control arm: case study in non-small cell lung cancer <https://www.focr.org/sites/default/files/pdf/SCA%20White%20Paper.pdf> (2018). Accessed December 22, 2020.
44. Hernán, M.A. & Robins, J.M. Estimating causal effects from epidemiological data. *J. Epidemiol. Community Health.* **60**, 578–586 (2006).
45. Greenland, S., Pearl, J. & Robins, J.M. Causal diagrams for epidemiologic research. *Epidemiology* **10**, 37–48 (1999).
46. Eichler, H.G. *et al*. "Threshold-crossing": a useful way to establish the counterfactual in clinical trials? *Clin. Pharmacol. Ther.* **100**, 699–712 (2016).
47. Eisenhauer, E.A. *et al*. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).
48. Greenland, S. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *Int. J. Epidemiol.* **30**, 1343–1350 (2001).