The Friends of Cancer Research Real-World Data Collaboration Pilot 2.0: Methodological Recommendations from Oncology Case Studies

Donna R. Rivera[1], Henry J. Henk[2], Elizabeth Garrett-Mayer[3], Jennifer B. Christian[4], Andrew J. Belli[5], Suanna S. Bruinooge[3], Janet L. Espirito[6], Connor Sweetnam[7], Monika A. Izano[7], Yanina Natanzon[8], Nicholas J. Robert[6], Mark S. Walker[8], Aaron B. Cohen[9], Marley Boyd[6], Lindsey Enewold[10], Eric Hansen[5], Rebecca Honnold[11], Lawrence Kushi[12], Pallavi S. Mishra Kalyani[1], Ruth Pe Benito[11], Lori C. Sakoda[12], Elad Sharon[10], Olga Tymejczyk[9], Emily Valice[12], Joseph Wagner[4], Laura Lasiter[13] and Jeff D. Allen[13,*]

The purpose of this study was to evaluate the potential collective opportunities and challenges of transforming real-world data (RWD) to real-world evidence for clinical effectiveness by focusing on aligning analytic definitions of oncology end points. Patients treated with a qualifying therapy for advanced non-small cell lung cancer in the frontline setting meeting broad eligibility criteria were included to reflect the real-world population. Although a trend toward improved outcomes in patients receiving PD-(L)1 therapy over standard chemotherapy was observed in RWD analyses, the magnitude and consistency of treatment effect was more heterogeneous than previously observed in controlled clinical trials. The study design and analysis process highlighted the identification of pertinent methodological issues and potential innovative approaches that could inform the development of high-quality RWD studies.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ The use of real-world evidence (RWE) in drug development is expanding and various applications are being investigated. Efforts to develop common methodological frameworks and align on key variable definitions are needed to support harmonized data collection and standards.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ A common collaborative research protocol was used across distinct real-world data (RWD) assets to assess the level of standardization capable across datasets and the utility of different real-world end points.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ Comparison of results highlights areas of concordance, suggesting that real-world time to next treatment line and real-world time to treatment discontinuation may be useful early clinical end points that may be used in prospective studies, although concerns regarding data missingness and potential biases are acknowledged.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ This study illustrates the power of multistakeholder collaboration to identify both the challenge and the importance of methodological rigor in RWD efforts to support generation of high-quality RWE.

The 21st Century Cures Act was enacted in 2016, to evaluate the health policy goals of trial design modernization and use of real-world evidence (RWE) creating numerous synergistic efforts.[1] These efforts aim to further understand ways in which RWE may complement or supplement randomized controlled trial data for regulatory purposes.[2–6] Evaluating applied examples of real-world data (RWD) and the generation of RWE are central to this effort. We build on the foundational work from the Duke-Margolis Center for Health Policy[2,3,6] and the National Academies of Science, Engineering, and Medicine (NASEM)[5] by highlighting

[1]US Food and Drug Administration, Bethesda, Maryland, USA; [2]OptumLabs, Cambridge, Massachusetts, USA; [3]Americal Society of Clinical Oncology, Alexandria, Virginia, USA; [4]IQVIA, Durham, North Carolina, USA; [5]COTA, Inc, Boston, Massachusetts, USA; [6]Ontada, Irving, Texas, USA; [7]Syapse, San Francisco, California, USA; [8]ConcertAI, Cambrdige, Massachusetts, USA; [9]Flatiron Health, New York, New York, USA; [10]National Cancer Institute, Bethesda, Maryland, USA; [11]Tempus, Chicago, Illinois, USA; [12]Kaiser Permanente, Oakland, California, USA; [13]Friends of Cancer Research, Washington, DC, USA. *Correspondence: Jeff D. Allen (jallen@focr.org)

Received July 15, 2021; accepted September 28, 2021. doi:10.1002/cpt.2453

considerations for study concept, design, and analysis, such as extracting key patient characteristic data and standardizing key variable definitions, with the goal of implementing a shared research protocol across distinct RWD assets. Friends of Cancer Research (*Friends*) collaborated with 10 data partners using oncology RWD from administrative claims, electronic health records (EHRs), prior authorization systems, and/or cancer registries, to conduct Pilot 2.0 evaluating outcomes for patients with advanced non-small cell lung cancer (aNSCLC) receiving systemic frontline therapies.[7]

Friends' Pilot 2.0 builds upon the results from Pilot 1.0, which included 6 data partners that evaluated the performance of real-world end points across multiple data sources.[4] These studies evaluated immunotherapy utilization for the treatment of aNSCLC to evaluate outcomes including overall survival (OS), which have previously indicated treatment benefit.[8,9] Furthermore, this subsequent study aimed to provide specific considerations in the development and design of RWD studies based on shared learnings arising from the observed variability among data sources. Methodological solutions were explored to address variability and enhance the alignment of target study populations for appropriate comparison of study outcomes. We discuss a potential strategy for standardization, including development of a common lexicon to describe and evaluate RWD quality, and share specific lessons learned from our experience implementing a common research protocol across varied RWD sources.

METHODS

Data partners and data sources

RWD partners that participated in this study represent data from a range of settings, including community oncology centers, academic medical centers, health systems, and integrated delivery system networks in the United States. The contributing partners included: ASCO CancerLinQ/ConcertAI,[10–13] Cancer Research Network,[14] COTA,[15] IQVIA,[16] Ontada (formerly McKesson),[17] SEER-Medicare,[18] Syapse,[19] and Tempus.[20] In addition, Flatiron Health, Mayo Clinic, OptumLabs, and Aetion also contributed to the early phase study design. Data curation included approaches that were unique to each participant based on availability, including natural language processing, artificial intelligence tools, technology enabled abstraction, and chart review. Common definitions were established (**Table 1**) and parallel analyses were performed by each group and summary results were submitted to FOCR.

Population

Each cohort selected patients with aNSCLC treated in the first line setting for advanced/metastatic disease with platinum doublet chemotherapy (PDC), PD-(L)1 monotherapy, or PD-(L)1 therapy in combination with platinum doublet chemotherapy (combination), as per the defined eligibility criteria (**Figure 1**). Patients were documented as having been physically present at a practice or as having had an encounter (defined as a physician visit, i.v. administration, or vitals documentation) in the database on at least 2 separate occasions on or after January 1, 2011, until data cutoff date (March 31, 2018). For the claims-based data source, patients were required to be enrolled on or after January 1, 2011, and before the data cutoff date (March 31, 2018). Determination of the end of follow-up (censor date) varied by participating institutions and was based on the most recent date for which complete information was available for the outcome of interest.

Population eligibility was limited to two primary factors that were known to be captured well across all data sources: (i) diagnosis: cancer type (aNSCLC) and (ii) treatment: documented receipt of a qualifying treatment regimen for advanced disease. Evidence of advanced disease was defined as stage IIIB, IIIC, or IV NSCLC at initial diagnosis or early stage (stages I, II, and IIIA) NSCLC with a recurrence or progression to metastatic disease (locally advanced or metastatic disease who had not received prior systemic therapy). The study maintained broad eligibility criteria, reflecting a real-world population. As such and due to varying levels of data availability between RWD sources, clinical characteristics often defined as eligibility criteria for clinical trials, such as organ function (renal and hepatic), PD-(L)1 status, and evidence of brain metastases, were not included.

Although histology was available from all data sources, it was included as a covariate in regression models but not as part of the inclusion/exclusion criteria due to sample size concerns. Adequate organ function is often considered in the use of PD-(L)1 therapy in routine clinical practice; however, laboratory values of organ function were not available from all data sources. Among the four data sources with available lab results, < 1% of patients were identified with severe hepatic or renal dysfunction, suggesting the treatment regimens studied were rarely used in patients with organ impairment and this exclusion, if applied, was expected to have little impact on study findings. Similarly, availability of PD-(L)1 status varied across the data sources and could not be included as a required covariate. Last, brain metastases may not be adequately captured in RWD sources, and lack of affirmative evidence of brain metastases was considered inadequate as a proxy for absence of brain metastases. Consequently, we adjusted for evidence of brain metastases but did not consider presence or absence of brain metastases in patient selection.

Frontline treatment

Treatment groups were defined based on exposure to PDC regimens (cisplatin/carboplatin, oxaliplatin, or nedaplatin with pemetrexed, paclitaxel, nab-paclitaxel, or gemcitabine), PD-(L)1 therapy (atezolizumab, nivolumab, or pembrolizumab), or combination therapy in the frontline setting. Treatment regimen was identified within each data source based on medication orders, medication administration records, medical claims, or infusion databases. Informed by expert clinical input, frontline regimen was defined as the first chemotherapy regimen given subsequent to the date of advanced diagnosis, and included all administered agents initiated within 30 days following the day of first infusion. All therapies were eligible for capture from the date of study initiation; however, it should be noted that approval of PD-(L)1 immunotherapy for aNSCLC did not occur until October 2015.

Study end points

The pilot included assessment of three end points: real-world overall survival (rwOS), real-world time to treatment discontinuation (of frontline regimen; rwTTD), and real-world time to next treatment line (rwTTNT). Overall survival (rwOS) was measured as the length of time from the date of first treatment administration in the frontline therapy regimen (index date) to the date of death or disenrollment (defined as the last known recorded clinical activity in structured data); however, completeness and validation of mortality data sources for rwOS varied across groups. End points that could be uniformly operationalized across RWD sources were chosen specifically for their capacity to convey important information associated with treatment benefit. Because disease progression is not uniformly defined nor captured in RWD sources, yet is clinically associated with regimen discontinuation or initiation of a new regimen or modality across therapeutic classes, rwTTD and rwTTNT were selected as measurable parameters to evaluate as end points instead of real-world progression-free survival, where rwTTD was defined as the length of time from the index date to the date of frontline treatment discontinuation.[10] There are notable limitations to interpretability of the TTD end point because the standard chemotherapy (PDC) regimens in the metastatic setting which are expected to continue for four to six cycles, whereas the use of PD-(L)1 therapy may continue indefinitely requiring

Table 1 Harmonized definitions employed in the pilot project

Term	Harmonized definition	Decision impact
Population		
Advanced NSCLC	All data sources had the ability to identify patients diagnosed with NSCLC. Evidence of advanced disease was defined as either stage IIIB, IIIC, or IV NSCLC or early-stage (stages I, II, and IIIA) NSCLC with a recurrence or progression at initial diagnosis.	Including patients diagnosed early stage (stages I, II, and IIIA) NSCLC with a recurrence or progression to advanced or metastatic status improved sample size for analysis but created a less homogeneous population of both newly diagnosed and previously treated (vs. patients newly diagnosed lung cancer).
Frontline	Patients were required to have no evidence of treatment in 180 days before the date of diagnosis and evidence of an eligible treatment within 120 days after diagnosis	Patients who have delays to treatment initiation would not be included.
Histologic subtype	Histology was not required for inclusion	Histology was not universally collected, although subanalysis feasible. Results reflected overall aNSCLC trends but were less specific to a histology subtype.
Eligibility criteria	The study population was not limited to those meeting eligibility criteria common for inclusion in a clinical trial (e.g., kidney function, performance status)	Data on organ function and performance status at or prior to treatment initiation was not often available or difficult to ascertain in RWD sources, although subanalysis was feasible. The population may be less like the RCT population(s).
Regimens		
Drugs	The following medications were included representing traditional chemotherapy or IO given after the date of diagnosis: cisplatin/carboplatin, oxaliplatin, or nedaplatin with pemetrexed, paclitaxel, nab-paclitaxel, or gemcitabine; atezolizumab, nivolumab, or pembrolizumab. Oral agents were not included.	Regimens are subject to misclassification, particularly in the doublet chemotherapy cohort. Patients starting on a PD-(L)1 should not be ALK or EFGR positive.
Frontline (first line regimen) assignment	Frontline regimen was defined as all administered agents received within 30 days following the day of first infusion.	Misclassification or omission of patients with delays to full treatment initiation in the first 30 days was possible. This would not impact the PD-(L)1 monotherapy cohort, as additional therapy would not be expected.
End points		
rwOS	Length of time from the date of treatment initiation to the date of death or end of follow-up; or end of study	Date of initiation may bias toward slightly shorter event times compared with clinical trials which can use date of randomization or enrollment instead. Missing events, on average, tend to make survival outcomes look better than in trials, especially if missingness is not independent of timing of death events.
rwTTNT	Length of time from the date of treatment initiation to the date of the next systemic treatment. When subsequent treatment is not received (e.g., continuing current treatment or disenrollment not due to confirmed death), patients were censored at their last known activity.	Missingness for subsequent treatment, including receiving treatment outside the system of capture is a limitation. This measure is also affected by the clinical guideline recommendations for administration of treatment cycles which can vary by regimen and has to be evaluated for comparability prior to the study to ensure appropriate interpretation.
rwTTD	Length of time from the date of treatment initiation to the date of patient treatment discontinuation the. The study treatment discontinuation date was defined as the last administration or noncancelled order of a drug contained within the regimen. Discontinuation was defined as having a subsequent systemic therapy after the initial regimen, having a gap of more than 120 days with no systemic therapy following the last administration, or having a date of death while on the initial regimen. Patients without a discontinuation were censored at the end of follow-up.	At the patient level, TTD is associated with PFS across therapeutic classes.[21]
rwTTP	Progression was omitted as claims-based algorithms are inadequate and among the EHRs progression events are not consistently captured in structured data. Unlike in clinical trials, there is not a uniform criterion (e.g., RECIST) in the off-protocol setting for determination of disease progression.	As TTP and PFS are accepted outcomes in clinical trials, comparison of these outcomes to randomized trials of similar regimens were limited by the data available.

(Continued)

Table 1 (Continued)

Term	Harmonized definition	Decision impact
Analysis		
Estimation	Kaplan-Meier estimation was used to describe distribution of end points for each dataset for each regimen, and for estimating key time points (e.g., 6-month, 12-month event rates) with confidence intervals.	
Comparisons	Proportional hazards regression, adjusting for prognostic factors available to all groups.	
Additional analyses	For OS, censor all events at 24 months and re-estimate HRs for treatment effect, adjusted for other prognostic variables.	

ALK, anaplastic lymphoma kinase; aNSCLC, advanced non-small cell lung cancer; EHR, electronic health record; HR, hazard ratio; IO, intra-osseous; NSCLC, non-small cell lung cancer; OS, overall survival; PFS, progression-free survival; RCT, randomized controlled trial; RECIST, Response Evaluation Criteria in Solid Tumors; rwOS, real-world overall survival; rwTTD, real-world time to treatment discontinuation; rwTTNT, real-world time to next treatment line; rwTTP, real-world time to treatment progression.

cautious interpretation of this end point in alignment with the clinical context for treatment. As a measure of regimen-specific treatment patterns, rwTTNT was defined as the time from the index date to the initiation of the subsequent regimen or date of death to assess changes in care. To avoid incorrectly identifying patients discontinuing treatment (e.g., leaving the health plan while still on treatment), the operational definition of rwTTD further required identification of patients whose event times were censored due to death or insufficient follow-up. Sufficient follow-up was defined as 120 days with no systemic therapy following the last administration of treatment in the frontline setting. Because mortality is more likely to be under-reported than over-reported in most widely available sources, rwOS, as presented in Kaplan-Meier curves in this study, likely overestimates the true rwOS distribution in this patient population (and could appear somewhat longer than corresponding data from clinical trials, which tend to have more complete follow-up requirements); nonetheless, rwOS is useful for evaluation in a real-world setting, especially in comparative evaluation of proxy end points in real-world studies, and was included as an end point in Pilot 2.0.

Data standardization and analysis

Collaborators jointly developed a common research protocol *a priori*, including definitions on patient selection criteria, key covariates, and outcomes, which were collected within a uniform reporting template

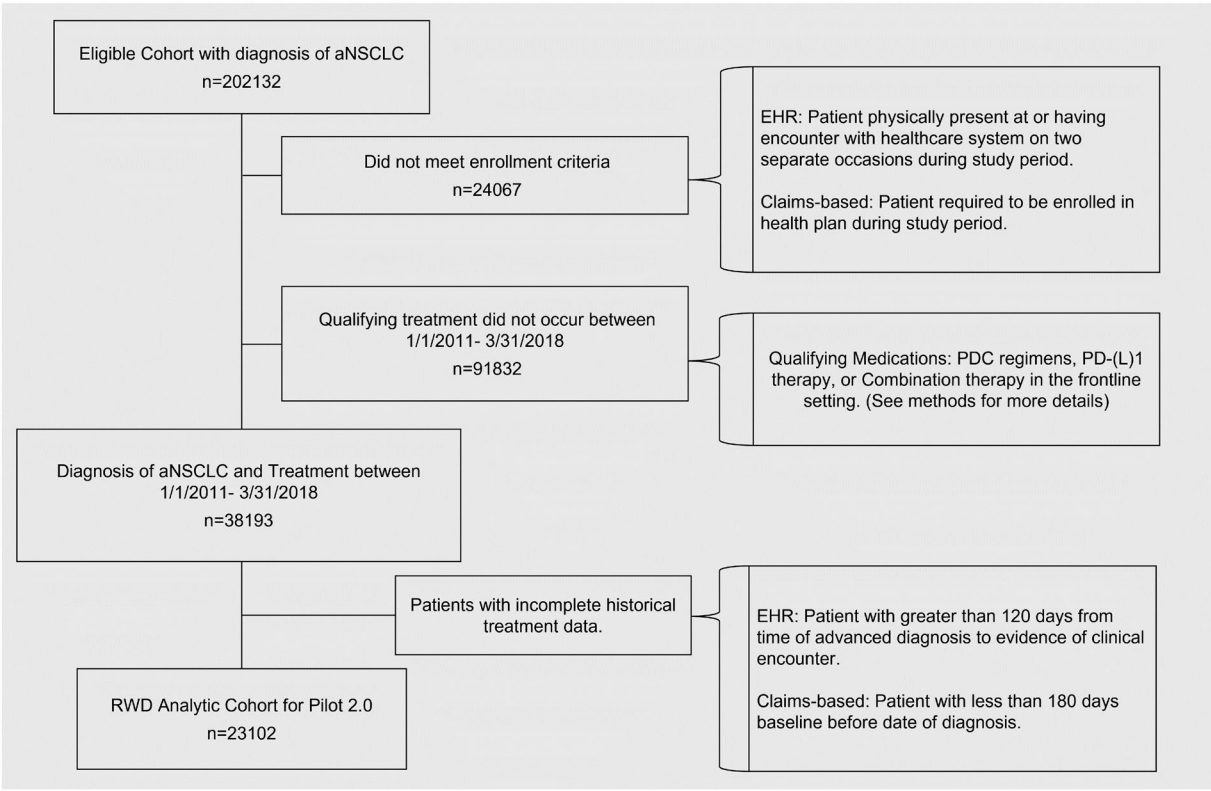

Figure 1 Cohort construction, including data from all data sources. aNSCLC, advanced non-small cell lung cancer; EHR, electronic health record; PDC, platinum doublet chemotherap; PD-(L)1, programmed cell death protein 1/programmed death-ligand 1; RWD, real-world data.

and accompanied by a detailed statistical analysis plan (**Supplementary Protocol**). Each RWD partner operationalized the common research protocol based on characteristics of each data source and technological feasibility, conducting analyses on their respective datasets individually, and reporting summary-level results via the uniform reporting template due to patient privacy, technical complexity, and proprietary nature of the datasets. It was deemed infeasible to aggregate or merge the data for analytic purposes and, instead, collaborators sought to standardize definitions and harmonize processes (**Table S1**). The development of each analytic dataset was subject to data availability and software programming accessible to each RWD partner. Each collaborator has established curation processes designed to evaluate the quality and completeness of their data. Thus, such research-ready databases may not reflect typical EHR data available or health insurance claims data that have not been subjected to such ongoing data curation.

Analytic methods were applied to account for several key sources of variation in the availability of follow-up data; specifically, impact of length of healthcare enrollment, year of treatment initiation, and therapy availability based on US Food and Drug Administration (FDA) approval dates. Kaplan-Meier estimation was used to describe the distribution of each end point (rwOS, rwTTD, and rwTTNT) for each regimen by RWD source. Presentation of the survival curves per regimen was important to assess differential follow-up across data source, and across regimens within sources, and to demonstrate censoring rates which varied across data sources. Proportional hazards regression was used to compare treatment arms for each end point, adjusting for prognostic factors reasonably available to all groups: status at diagnosis (advanced at diagnosis vs. early stage and progressed to aNSCLC), stage, age, year of treatment initiation, gender, race, histology, smoking history, PD-(L)1 expression status, Eastern Cooperative Oncology Group (ECOG) performance status, and time to treatment initiation from diagnosis. All prognostic factors were included as nominal categorical covariates, with continuous variables (e.g., age) converted to categorical scales. To address missingness and avoid excluding patients who had missing values for any of the prognostic factors, each covariate included a "missing/unknown" category. Hazard ratios (HRs) from regression models comparing PD-(L)1 to PDC and combination to PDC are presented with 95% confidence intervals (CIs).

The study included patients initiating treatment for aNSCLC in 2011, whereas approval of PD-(L)1 immunotherapy for aNSCLC did not occur until October 2015, skewing the length of available follow-up data toward the PDC cohort. Uptake of PD-(L)1 and combination increased during 2016 and combination during 2017–2018, respectively, limiting follow-up, comparisons, and inferences related to this cohort to a 2-year period. A sensitivity analysis was conducted censoring all patients without an event at 24 months and re-estimating HRs for treatment effect to assess the impact of differential follow-up among the different treatment cohorts (results not shown).

RESULTS

Clinical and demographic characteristics were similarly distributed within treatment groups for each RWD source (**Figure 2**). Geographic coverage varied by data source. There are notable differences in missingness of certain variables across data sources (e.g., smoking and performance status; **Figure 2**). Overall utilization of PD-(L)1 and combination regimens increased over the study period, with the earliest use of PD-(L)1 therapy starting in 2015.

Median rwOS ranged from 10–17 months for PDC across data groups (**Figure 3a**) and was 12–18 months in the PD-(L)1 groups (**Figure 3b**). Kaplan-Meier curves for rwOS, rwTTD, and

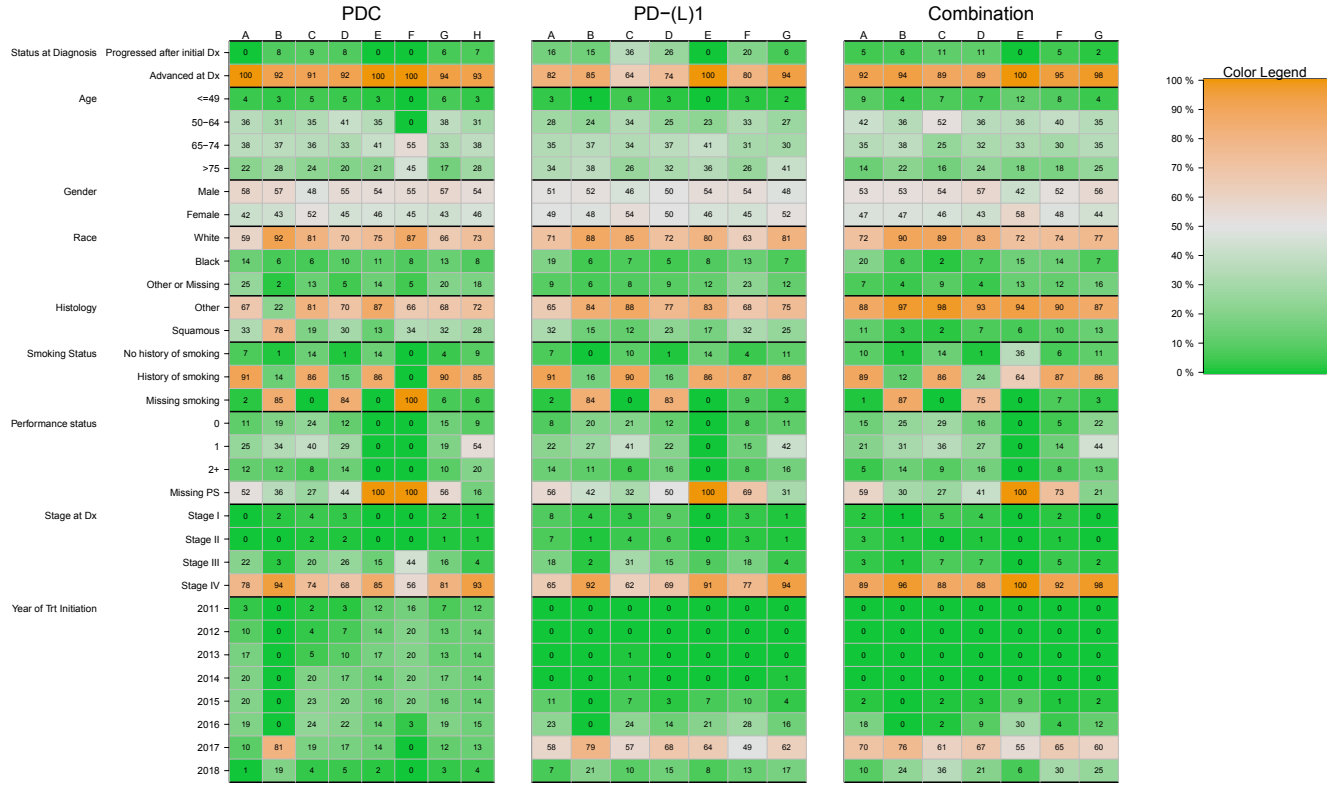

Figure 2 Characteristics of treatment groups within participating data sources. Numbers in table represent percent of patients in each category. Coloring ranges from bright green (0%) to bright orange (100%) to highlight areas of differences across data sources for the same treatment.

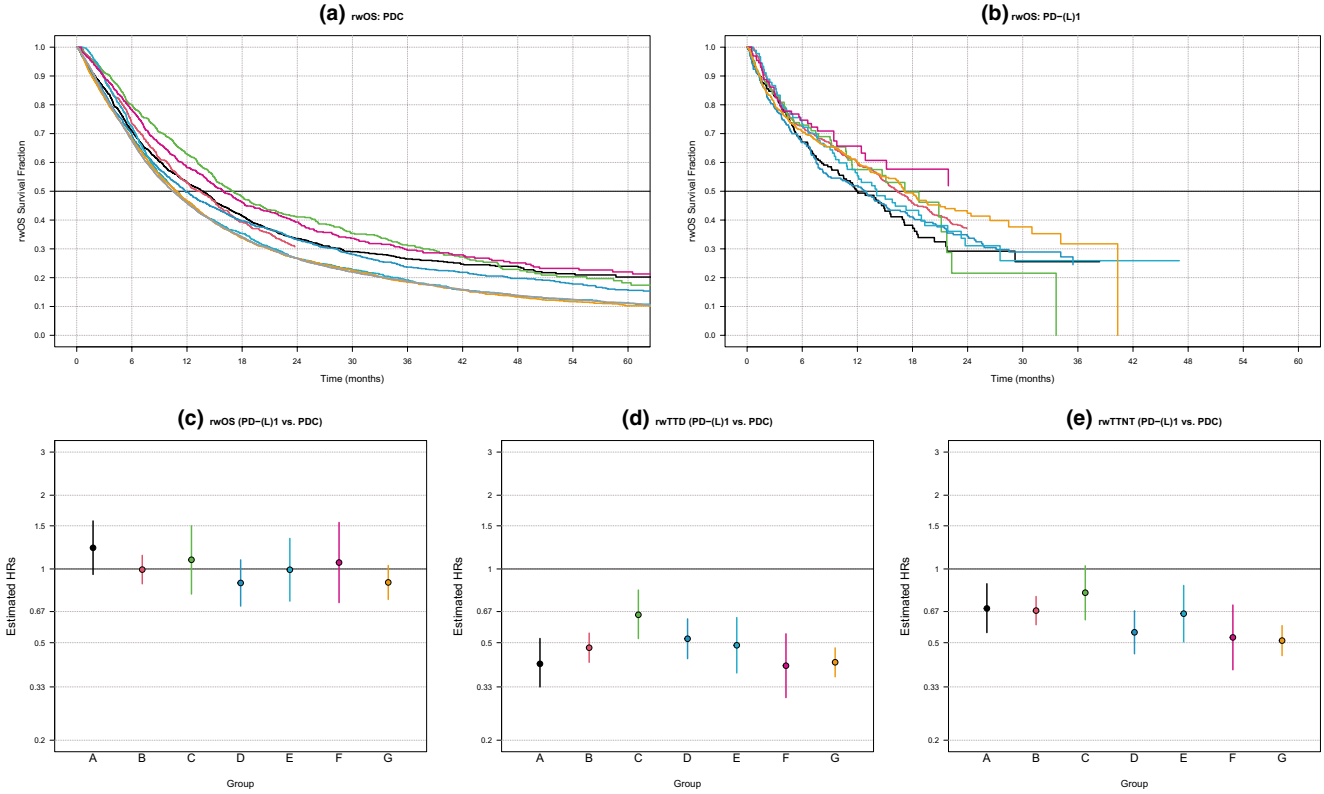

Figure 3 Kaplan-Meier curves for overall survival (OR) for (**a**) PDC and (**b**) PD-(L)1; hazard ratios with 95% confidence intervals (vertical bars) comparing PD-(L)1 to PDC for (**c**) OS, (**d**) time to treatment discontinuation, and (**e**) time to next treatment (TTNT). PDC, platinum doublet chemotherapy; rwOS, real-world overall survival; rwTTD, real-world time to treatment discontinuation; rwTTNT, real-world time to next treatment.

rwTTNT for all treatment groups are provided in **Figures S1**, **S2** and **S3**. The 1- and 5-year rwOS estimates for PDC were within a similar range across datasets. HRs (with 95% CIs) in **Figure 3c** comparing rwOS for PD-(L)1 vs. PDC, adjusted for a common set of covariates, suggest no evidence of association: HRs range from 0.88 to 1.22 with all 95% CIs overlapping 1. The direction of the association varied among data partners, with 3 having HR estimates greater than 1 (1.06, 1.09, and 1.22), 2 with HR estimates less than 1 (0.88 and 0.88), and 2 with HR estimates at almost exactly 1 (0.99 and 0.99). Although there was consistency in rwOS curves across data sources within each treatment group for PD-(L)1 and for PDC through ~ 6 months, there was more variability in rwOS curves in the 0-to-6-month time period in the combination rwOS curves (**Figure S1**). This may be due to smaller sample sizes in the combination cohorts, leading to more imprecise estimates. For both rwTTD and rwTTNT, HRs were less than 1 (in all but one case; **Figure 3e**, **group C**), demonstrating that patients on PD-(L)1 had longer times on treatment than patients who received frontline PDC in these populations studied. There were differences observed in rwTTD, with HR (range, 0.40 to 0.65; **Figure 3d**); however, these results should be considered in the context of certain therapeutic regimens having a set number of cycles of therapy prior to discontinuation, particularly in the PDC cohort. Associations were also observed for rwTTNT, with smaller effect sizes, with HR (range, 0.51 to 0.80; **Figure 3e**). Similar results were observed when comparing combination to PDC (**Figure S4**). The evaluation of results and conduct of this RWD study led to a set of methodological best practices when designing a RWD study across a multistakeholder group (**Table 2**). Discussion of the importance of the considerations of confounding by key prognostic factors is discussed in Supplementary Materials (**Figure S5**).

DISCUSSION AND RECOMMENDATIONS

The Friends' RWE Pilot 2.0 was a collaborative effort among participating data partners building upon prior work conducted under Pilot 1.0.[4] The updated collaboration further evaluated the performance of real-world end points across RWD sources in answering a common clinical question, specifically on outcomes among patients with NSCLC who received treatment (PDC, PD-(L)1, or combination) in the first-line advanced or metastatic setting. Through a collaborative common protocol, all RWD partners used the following process steps and engaged in weekly communication for RWD evaluation: (i) common shared protocol and statistical analysis plan (SAP), (ii) standardizing definitions across datasets, (iii) variance in methodological approaches (acceptable variance), (iv) data quality and sensitivity analysis, and (v) transparency by reporting limitations. The study showed broadly similar patterns of outcomes in which the distribution of rwOS was consistent across treatment cohorts among the patient cohorts with similar characteristics (**Figure 2**), but PD-(L)1 containing regimens had longer TTD and TTNT than PDC, demonstrated by estimated HRs in **Figure 3** and **Figure S4**. However, there was

Table 2 Recommendations from the RWE Pilot 2.0 for developing a common RWD protocol to achieve consistency and increase reproducibility using a format that minimizes ambiguity or subjectivity in interpretation of definitions or analysis approaches

Recommendation/ Sub-recommendation	Description
Defining the eligibility criteria	Shared variables that are commonly available across data sources should be used for defining patient inclusion in the study. In the RWE Pilot 2.0, cancer diagnosis (including stage and cancer type) and treatments receipt (platinum doublet and immunotherapy) were the primary criteria. Given that the goal was to make real-world inferences, the eligibilities were based on the population of interest for generalizability. For example, if the goal was to assess treatment differences in patients with advanced age, the age range would be limited to adequately address that question in this study.
Collaborative common RWD protocol	The collaborative protocol should determine a list of core common required data elements, common variables available in a variety of formats that require translation (e.g., age groups; gender and race categories) should be described, definitions (e.g., exposure and end points) should be included, and standardized reporting formats should be agreed on prior to study initiation. Include a standard reporting template complete with table and figure drafts to create understanding around the intended results to be generated.
Define core common key data elements	Establish a core set of data elements with standard definitions enables greater comparability. Variables may have varying levels of availability in RWD, and their relevance for inclusion as a required variable depends on the relation to the study question. Structured data such as age and sex, are minimal common data elements that are typically readily available across independent data sources and requisite for analysis. However, other data elements demand thoughtful consideration and transparency such as (ii) variables available in different formats (e.g., PD-L1 biomarker +/− indicator vs. expression), (ii) variables requiring derivation (e.g., ICD codes vs. laboratory values in the definition of reduced organ function), or (iii) variables requiring extraction from unstructured data (e.g., status of advanced at initial diagnosis vs. progression after initial diagnosis).
Align clinical variables and laboratory values	Key clinical and analytic variables should be identified and aligned as needed, and it should be determined whether strict variable definitions are required for inclusion criteria or if variations are acceptable. Variance in measurement can lead to subsequent impact on outcome calculations. For example, kidney function or genomic testing may be extracted from structured or unstructured data, where a source could have data ranging from the actual lab values to markers of function (e.g., laboratory tests for organ function, CrCl, ICD-9/10 indicating dysfunction) or indicators of testing to specific testing results (e.g., PD-(L) test completed to expression percentage). In areas where variation is accepted, the use of sensitivity analyses to examine variance is useful to guide inappropriate interpretation. Implement a well-developed common protocol for all RWD studies *a priori* to ensure internal and external replication.
Data quality assessment	Development of a template for quantitative evaluation of data distributions, quality, and missingness may provide a quantitative approach to understanding data availability and missingness for improved interpretation. However careful evaluation by a representative team that has deep knowledge of the data curation, extraction, and provenance is necessary. The use of quality indicators for data or consensus on problematic missingness for key covariates may inform the study design.
End point selection	Commonly used end points in clinical trials may not be practical or replicable in RWD. As an example, rwTTP and rwPFS were not included in Pilot 2.0. Challenges with measuring rwTTP and rwPFS exist: claims-based algorithms are limited, relying on proxy measures for progression and consensus definitions among EHRs data sources were prohibitively difficult to establish because of differences in capture and reporting. While uniform criterion (e.g., RECIST) allow protocol directed establishment of progression in clinical trials, progression outcomes are not consistently captured in RWD as there is currently a lacking capability in the off-protocol setting for determination of disease progression. Additional endpoints, rwTTNT and rwTTD, are more readily accessible in RWD. While survival outcomes (rwOS) are easier to define and measure in most RWD sources, sources are often missing mortality information on a large fraction of patients, which affects estimation of rwOS parameters (e.g., median rwOS) and substantially limits interpretation, while incurring additional biases due to missing data. Linking to additional data sources which include more complete mortality data could improve end point ascertainment and should be done if feasible to make estimates based on rwOS more accurate and evaluable to other studies, such as clinical trials.
Defining event times and censoring	When evaluating endpoints, there is a need to it may be most reasonable merge clinical applicability with analytical feasibility. For example, in defining rwTTD, groups had to align on the appropriate time period that would equate with without no treatment receipt to be considered a discontinuation. An additional step in the process would be evaluating the potential to share software code between groups for replicability and additional validation.
Statistical analysis plan	SAP must be written comprehensively with sufficient detail to reduce the risk of deviations in methods used and characterizations of variables in models or tests. In conjunction with the SAP, it is instrumental that the protocol includes table and figure templates to ensure that all groups have the same understanding of the intended results to be generated, and the models required to reduce variance in interpretation. Developing tables within the shared research protocol allowed groups to consider subtle differences in modeling that would not have arisen without having developed them in advance.

(Continued)

Table 2 (Continued)

Recommendation/ Sub-recommendation	Description
Addressing missing data and potential biases	Approaches for quantifying and accounting for missing data in analyses should be considered in the protocol to maintain study integrity while minimizing biases in the interpretation of results.[24] Data missingness should be evaluated by a team that has deep knowledge of the data curation, extraction, and provenance. Imputation should be carefully considered, given the potential for missingness of variables to be related to patient outcomes (i.e., informative missingness) in RWD; choices such as imputation of the data or use of the missing category in modeling have implications for study analyses and inferences. Additionally, use of bias quantification approaches may be useful in appropriate interpretation of the results and understanding study limitations, which in RWD are often limitations of the underlying data.
Assess sample size	Because the number of patients in RWD sources is often based on retrospective data availability, study planning for RWD studies may not consider sample size and the power to detect clinically relevant effects. Even so, it is important to ensure that the sample size is sufficiently large to be able to derive meaningful inferences. If the study is underpowered, modeling may be infeasible or hypothesis tests can tend to find "insignificant" findings with wide confidence intervals, leading to potentially misleading results. In contrast, if the RWD source provides a very large sample, the study may be overpowered and there will be a tendency to over-interpret statistically significant findings. Statistically significant *P* values do not necessarily imply clinical significance. In that case, interpretation of results could focus effect estimates with their confidence intervals (or similar quantities), and not necessarily alone on *P* values.
Cautious inference	Even with careful attention to adjustment for population differences, there are inherent selection bias and unmeasured confounding as well as cohort effects that may not be able to be accounted for in a study; these limitations of RWD need to be appropriately addressed in the interpretation of results, inferences, and conclusions of RWE studies. In our study, while there were no obvious differences in the patient characteristics included in Pilot 2.0 across treatment cohorts, the clinical standard of care was likely to differ for the PDC population before and after FDA approval for PD-(L)1 therapies. Similarly, comparisons of results from RWE studies to results from clinical trials need to be cautious given underlying differences in patients treated in clinical trials vs. those in available in RWD sources; these differences are expected due to limited adult clinical trial participation in patients with cancer (3–5%) and strict trial eligibility criteria. This is a strength of RWD in allowing expansion of eligibility criteria to better understand use in a real-world population which is, in turn, a limitation in comparative efforts due to the aforementioned selection bias.
Diverse Multidisciplinary Research Team	Perhaps the most pivotal part of the process in an RWD study is developing a multidisciplinary team, including clinicians, biostatisticians, epidemiologists, and data scientists, to ensure that studies are clinically relevant with appropriate methods utilized to optimally account for potential biases arising from the observational nature of RWD. Teams are encouraged to include patient stakeholders and diverse representation in the conversation, as this is most effectively accomplished as a team science effort.

CrCl, creatinine clearance; EHR, electronic health record; ICD, International Classification of Disease; RECIST, Response Evaluation Criteria in Solid Tumors; RWD, real-world data; RWE, real-world evidence; rwOS, real-world overall survival; rwPFS, real-world progression-free survival; rwTTD, real-world time to treatment discontinuation; rwTTNT, real-world time to next treatment line; rwTTP, real-world time to treatment progression.

notable variability in parameter estimates across data sources, differences in the level of missingness of certain variables of interest to the analysis; and, therefore, differences in subgroup evaluations, as shown in **Figure 3**.

This study demonstrated a successful collaboration aimed at examining research methodologies to provide approaches to measure treatment effects for patients treated in real-world settings and highlights continuing challenges regarding how to best use RWE. For example, different sources of RWD may arrive at similar conclusions regarding relative effectiveness of therapies or heterogeneity may emerge that is not sufficiently able to be overcome or interpreted. Compared with other end points evaluated, rwTTNT showed greater consistency of findings across sources (**Figure S2**). However, all the outcomes reported here were more consistent across data sources for the patients treated with PDC than for those treated with a PD-(L)1 agent or combination, seemingly because of the greater number of chemotherapy recipients (improving precision of estimates for PDC parameters). This suggests that rwTTNT may be less susceptible to the impact of the data variations that exist among RWD sources than other end points, but also highlights the importance of sample size in the stability of parameter estimates. Additionally, whether rwTTNT could be appropriate as an end point in a comparative RWD analysis for a regulatory objective would require additional considerations as this end point does not strictly measure efficacy; further validation would be necessary as it has not been evaluated in a clinical trial setting. However, it could be considered alongside other measures in a pragmatic prospective design. Additional development and validation of these end points is needed, including further exploration around the guidance for clinical use cases for real-world end points as well as ways in which they can be constructed to ensure appropriate interpretability of findings. Although RWD studies may provide an opportunity for increased generalizability and access to expanded populations, study-specific sample size calculations are still necessary to inform study feasibility.

Widely used end points in traditional cancer clinical trials may not be practical or replicable across diverse sources of RWD and pose clear challenges to implementation, particularly when using RWD to construct an external control arm. For example, RWD sources face current challenges with measuring progression-based endpoints (e.g., real-world progression-free survival and

real-world time to treatment progression) including lesser accessibility in claims sources where measurement algorithms are not well-established. Disease progression determination tends to be subjective and may not be uniformly defined across or within and EHR-based data sources and is a future area of investigation for the RWE pilot projects. In contrast, OS is objectively defined, and represents the least variable real-world end point. The use of rwOS, particularly in prospective randomized pragmatic trials, represents an opportunity for use of RWD, particularly where the length and size of such a study would be considered impractical in the traditional clinical trial setting. Despite this potential advantage, it is documented that RWD may be missing mortality information on a large set of patients, which can limit the utility of rwOS in these instances especially in a regulatory context. Understanding the mechanism of missingness (and whether it is random or nonrandom), and potentially incorporating mortality data from external sources are considerations for inferences related to rwOS.[22]

Pilot 2.0 illustrates the importance of considering the potential for selection biases present in RWD, and furthermore in the evaluation of these considerations across diverse data sources. The conditions which cause patient information to be present within a given data source may not be at random, could be associated with the outcome or exposure of interest, and may also be subject to systematic information biases.[23] The likelihood that patient information may be present in a particular source may depend on the practice, treating physician, geography, age, employment, income level, social determinants of health, legal residency status of patients, or other ascertainment practices by the data partner. These factors may be prognostic and therefore associated with the outcome of interest. The evaluation of these types of systematic biases, including selection and information bias, required a multidisciplinary team evaluation to ensure the clinical, statistical, and epidemiological factors are evaluated adequately. Future research to quantify the impact of these biases would improve the ability to interpret the impact of study variance (in between cohorts and among data sources) for patient, clinical, and regulatory decision making. This research also shows the importance of intentionally considering the clinical perspective of how care delivery and treatment may have changed over time (including evaluation of time varying confounding), as well as how the recency of the data and the duration of follow-up may affect the study results regarding treatment patterns described by the data. The heterogeneity present across the Pilot 2.0 data sources also exemplifies the challenges in interpretation of RWD within the context of evidence from traditional clinical trials, especially in any direct comparative or emulative efforts. The real-world sample is less likely to be highly selected with increased comorbidities and increased diversity. Thus, RWD studies may show overall treatment effects that are more modest than those reported in trials; however, they likely could be more representative of the real-world experience of patients. This may provide increased generalizability, especially to the intended treatment population. Establishing population representativeness, including being nationally representative, was beyond the objectives of this study. Acknowledging the benefit of improved representativeness, lack of randomization in retrospective RWD analyses makes causal inference on comparisons challenging and thus differential benefits or harms seen when comparing nonrandomized RWD cohorts should be considered hypothesis generating at present.

Future research on methods for standardization of an approach to categorize data and establish objective measures of data quality that incorporate pertinent RWD assessments is needed. Establishing objective measures of data quality that incorporate assessments of longitudinality, temporality, missingness, and representativeness, perhaps benchmarked against key features of established datasets, would represent an essential advance. This study was limited by inherent factors of retrospective observational research, and by factors unique to the collaborative nature of this effort. First, although data sources were independent, it is unknown to what extent patients are represented in more than one source because of limitations around patient level data access as well as the inability to do any type of matching (comparability). To the extent this occurred, there could be duplication and sources would tend to appear more similar. Second, the data partners conducted their analyses independently, albeit following a carefully developed and detailed analysis plan. Nevertheless, different software packages were used which may have allowed for use of slightly different methods in areas not specifically governed by the analysis plan. Third, there was substantial missingness in certain data types across the data sources. This is likely to have influenced OS estimates for these sources. Additionally, information regarding oral agents was not included and the analysis was not able to account for receipt of certain targeted therapies, which may have been more common in the PDC cohort. Unobserved factors may have influenced receipt of specific treatments. Last, control of potential confounding in this study was limited by the need to implement a uniform analysis across data partners. This approach allowed for comparability of findings across data sources; however, patient characteristics that could have been included as potential confounders in analyses within individual data sources were not included in adjusted analyses in the interest of the broader research goal.

Lessons from the experience of Pilot 2.0 are presented as recommendations for future work in **Table 2**. Friends and the data partners have shown that a diverse group of research enterprises can collaborate effectively to advance the use of oncology RWE. Comparison of results highlights areas of concordance, suggesting that rwTTNT and rwTTD may be useful early clinical end points that may be used in prospective studies, although concerns regarding data missingness and potential biases are acknowledged. In summary, the study illustrates the power of multistakeholder collaboration to identify both the challenge and the importance of methodological rigor in RWD efforts.

SUPPORTING INFORMATION

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

FUNDING

No funding was received for this work.

CONFLICT OF INTEREST

H.J.H. is an employee of OptumLabs. E.G.-M. has received consulting/advisory fees from Deciphera and TYME. J.B.C. owns stock in IQVIA. A.J.B. and E.H. have ownership in COTA, Inc. J.L.E. owns stock in McKesson. C.S. and M.I. own stock in Syapse. Y.N. owns stock in Syapse and ConcertAI and received travel/accommodations by Sypase. N.J.R. holds a leadership position in McKesson; stock/ownership in Johnson & Johnson, McKesson, and Oncolytics Biotech; holds honoraria with Bristol-Myers Squibb and Roche; and consulting roles for ADVI, Boehringer Ingelheim, Bristol-Myers Squibb, and New Century Health. M.S.W. is an employee of ConcertAI. A.B.C. owns equity in Flatiron Health, a subsidiary of Roche and stock in Roche. R.H. is an employee of Tempus. L.K., L.C.S., and E.V. are employees of Kaiser Permanente. R.P.B. is an employee of Tempus. O.T. owns stock in Roche. J.W. owns stock in IQVIA and Merck. All other authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

D.R.R., H.J.H., E.G.M., J.B.C., A.J.B., S.S.B., J.L.E., C.S., M.A.I., Y.N., N.J.R., M.S.W., A.B.C., M.B., L.E., L.K., P.S.M.K., R.P.B., L.C.S., E.S., O.T., E.V., L.L., and J.D.A. wrote the manuscript. D.R.R., H.J.H., E.G.M., J.B.C., A.J.B., J.L.E., M.A.I., Y.N., N.J.R., A.B.C., M.B., L.E., L.K., P.S.M.K., R.P.B., E.S., O.T., L.L., and J.D.A. designed the research. H.J.H., C.S., E.H., R.H., and J.W. performed the research. D.R.R., H.J.H., E.G.M., A.J.B., J.L.E., M.A.I., Y.N., N.J.R., A.B.C., L.E., R.H., O.T., J.W., and L.L. analyzed the data.

DISCLAIMER

The views represented in this paper represent the individual authors and should not be interpreted as representing any official HHS views or policy.

1. US Food and Drug Administration. *Framework for FDA's Real-World Evidence Program* (US Department of Health & Human Services, Washington, DC, 2018).
2. Daniel, G. *et al*. *Characterizing RWD Quality and Relevancy for Regulatory Purposes* (Duke-Margolis Center for Health Policy, Washington, DC, 2018)
3. Mahendraratnam, N. *et al*. *Determining Real-World Data's Fitness for Use and the Role of Reliability* (Duke-Margolis Center for Health Policy, Washington, DC, 2019).
4. Stewart, M. *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.* **3**, 1–15 (2019).
5. National Academies of Sciences, Engineering, and Medicine 2019. *Examining the Impact of Real-World Evidence on Medical Product Development: Proceedings of a Workshop Series* (The National Academies Press, Washington, DC, 2019).
6. Mercon, K. *et al*. *A Roadmap for Developing Study Endpoints in Real-World Settings* (Duke-Margolis Center for Health Policy, Washington, DC, 2020).
7. Bin, M. & Ong, H. Real-world evidence at glance: How a collaboration of "frenemies" produced common definitions for real-world endpoints. *Cancer Lett.* **45**, 5–12 (2019).
8. Howlader, N. *et al*. The effect of advances in lung-cancer treatment on population mortality. *N. Engl. J. Med.* **383**, 640–649 (2020).
9. Marur, S. *et al*. FDA analyses of survival in older adults with metastatic non-small cell lung cancer in controlled trials of PD-1/PD-L1 blocking antibodies. *Semin. Oncol.* **45**, 220–225 (2018).
10. ASCO CancerLinQ. LLC: Lung Data (2011–3/31/2018) <www.cancerlinq.org/solutions/researchers>.
11. Schilsky, R.L. *et al*. Building a rapid learning health care system for oncology: the regulatory framework of Cancer-LinQ. *J. Clin. Oncol.* **32**, 2373–2379 (2014).
12. Shah, A. *et al*. Building a rapid learning health care system for oncology: why CancerLinQ collects identifiable health information to achieve its vision. *J. Clin. Oncol.* **34**, 756–763 (2016).
13. ConcertAI Patient360 <https://www.concertai.com/data-products/> (2021).
14. National Cancer Institute. Cancer Research Network <https://healthcaredelivery.cancer.gov/crn/>.
15. COTA, Inc. COTA Overview <https://cotahealthcare.com/>.
16. IQVIA. Harness the power of Real World Data <https://www.IQVIA.com/solutions/real-world-evidence/real-world-data-and-insights>.
17. Ontada. The US Oncology Network <https://www.usoncology.com> (2020).
18. Warren, J.L. *et al*. Overview of the SEER-Medicare data: content, research applications, and generalizability to the United States elderly population. *Med. Care* **200240**(Suppl. 8):IV-3–IV-18.
19. Lerman, M.H. *et al*. Validation of a mortality composite score in the real-world setting: overcoming source-specific disparities and biases. *JCO Clin. Cancer Inform.* **5**, 401–413 (2021).
20. Fernandes, L.E. *et al*. Real-world evidence of diagnostic testing and treatment patterns in US patients with breast cancer with implications for treatment biomarkers from RNA sequencing data. *Clin. Breast Cancer* 1526–8209 (2020).
21. Blumenthal, G.M. *et al*. Analysis of time-to-treatment discontinuation of targeted therapy, immunotherapy, and chemotherapy in clinical trials of patients with non-small-cell lung cancer. *Ann. Oncol.* **30**, 830–838 (2019).
22. Curtis, M.D. *et al*. Development and validation of a high-quality composite real-world mortality endpoint. *Health Serv. Res.* **53**, 4460–4476 (2018).
23. Tripepi, G., Jager, K.J., Dekker, F.W. & Zoccali, C. Selection bias and information bias in clinical research. *Nephron Clin. Pract.* **115**, c94–c99 (2010).
24. Lash, T.L. *et al*. Good practices for quantitative bias analysis. *Int. J. Epidemiol.* **43**, 1969–1985 (2014).