FRIENDS of CANCER RESEARCH

A FRIENDS OF CANCER RESEARCH WHITE PAPER

Considerations for Use of Real-World Evidence in Oncology

LESSONS LEARNED FROM FRIENDS OF CANCER RESEARCH COLLABORATIONS

Objectives

- Discuss methods and considerations for extracting data on patient characteristics
- Standardizing definitions/methodology across multiple RW datasets with the intent of aligning to similar patient populations
- Describe opportunities and potential problems in allowing flexibility in definitions
- Processes for assembling "fit-for-purpose" real-world datasets

Clinical trials, from Phase I dose-finding and safety trials to Phase III randomized trials examining efficacy, form the backbone of the drug development pipeline and inform regulatory approvals. While the centrality of clinical trials remains, there has been increasing interest in the potential contributions of real-world evidence (RWE) that results from analyses of real-world data (RWD). RWD refers to information that is collected during standard clinical care or health care billing, such as in electronic health records (EHR) or health insurance claims data, and can be leveraged for research and analytic purposes. The resulting evidence generated, called RWE, can reflect broader, more diverse patient populations than are typically included in traditional clinical trials and can be applied across multiple use cases, including to answer timely clinical questions, assess endpoints measures, perform comparative effectiveness research, and study long-term drug safety. Still, challenges remain on how to realize the full potential of RWE to support clinical research, drug development, and regulatory decision-making. Standardized variable definitions within datasets, harmonization across datasets, and application of appropriate analytical methods remain important considerations and challenges.

Recent implementation of legislative and regulatory policies focused on RWE, such as the 21st Century Cures Act, Prescription Drug User Fee Act (PDUFA) Reauthorization of 2017, and FDA Framework on Real-World Evidence, highlight the interest in using RWE applications across the drug development life cycle. Building trust in routine use of RWE for regulatory decisions will require a firm understanding of the question being asked, underlying data across real-world datasets, including the various sources of available data, their strengths and limitations, and the implications for observed endpoints. Well-validated endpoints must also be assessed as real-world endpoints to support the acceptance of RWE. Multi-

Friends of Cancer Research extends our thanks to the project partner organizations and working group members

PARTNER ORGANIZATIONS

ASCO

CHU Nantes

ConcertAI

COTA

Flatiron Health

Health Data Insight CIC

IQVIA

Kaiser Permanente

McKesson

National Institutes of Health

Optum Labs & Mayo Clinic

Owkin

Syapse

Tempus

U.S. Food and Drug Administration

WORKING GROUP MEMBERS

ASCO
Suanna S. Bruinooge
Elizabeth Garrett-Mayer

CHU Nantes
Pr Brigitte Dréno
Cécile Frénard
Romain Goussault
Emilie Varey

ConcertAI
Whitney Rhodes
Mark S. Walker

COTA
Andrew Belli
Eric Hansen
Keshava Dilwali
C.K. Wang

Flatiron Health
Shrujal Baxi
Aaron B. Cohen
Nicole Mahoney
Erik Rasmussen
Olga Tymejczyk
Aracelis Z. Torres

IQVIA
Jennifer Christian
Nancy Dreyer
Christen Gray
Pia Horvat
Adam Reich
Joe Wagner
Xinyan Yu

Kaiser Permanente
Lawrence H. Kushi
Lori C. Sakoda
Emily Valice

McKesson
Marley Boyd
Janet Espirito
Nicholas Robert

NIH
Eric Boyd (contractor with Information Management Services Inc., Calverton.)
Lindsey Enewold
Donna Rivera
Elad Sharon

OptumLabs
Henry Henk

Owkin
Etienne Bendjebbar
Maxime He
Anna Huyghues-Despointes
Anne-Laure Moisson
Radha Patel

Syapse
Monika Izano
Yanina Natanzon
Connor Sweetnam

Tempus
Ruth Pe Benito
Rebecca Honnold

stakeholder collaboration is necessary to develop robust recommendations to maximize the quality and utility of RWD analyses. This includes selecting datasets and sources that are appropriate and fit-for-purpose to address the question being addressed, as well as the subsequent evidence generation that is needed in support of oncology research, including drug development.

Informed by several pilot projects leveraging a common protocol (established in the RWE 1.0 Pilot Project and expanded upon in subsequent pilots, described below) among multiple real-world data partners, US and international populations, and oncology-specific disease settings, Friends of Cancer Research (*Friends*) and collaborators identified implications of dataset specifics and patient characteristics on real-world endpoints and recommendations for developing a RWE framework to encourage and guide future RWE studies that leverage multiple data sources to answer a single question through a harmonized protocol.

Friends of Cancer Research Real-World Evidence Pilots

RWE 1.0 Pilot Project

The initial *Friends* RWE Pilot, 1.0, brought together six data partners to evaluate the performance of real-world endpoints across multiple data sources by focusing on a common clinical question: What endpoints for advanced non-small cell lung cancer (aNSCLC) patients treated with immune checkpoint inhibitors can be evaluated and compared across all of these data sources? To answer this question, the RWE Pilot 1.0 members aligned on a framework of necessary data elements, characteristics, definitions for real-world (rw) endpoints, based on data availability in electronic health record (EHR) and claims systems. The preliminary goal was to evaluate whether the various datasets included in this study could achieve similar results when measuring treatment effect using a common framework. The protocol developed through the RWE Pilot 1.0 served as the basis for several additional pilot projects aimed at (1) identifying minimum data quality and reporting standards to aid the interpretation of individual RWD studies, and comparisons across studies performed using different RWD sources, (2) evaluating the ability to estimate and compare effectiveness endpoints for different therapies across the data sources, and (3) adapting this framework for evaluation of rw endpoints in the context of a specific research question. Results from RWE Pilot 1.0 showed that similar patient populations could be extracted across datasets with differing underlying data sources using aligned baseline characteristic definitions and a harmonized protocol, and that certain rw endpoints, time-to-treatment discontinuation (rwTTD), were correlated with rw overall survival (rwOS).

RWE Pilot 1.0 was then extended to other data sources and disease settings in an effort to further examine the generalizability of the findings and the framework.

RWE Pilot 1.0: RWE Framework in the United Kingdom Cancer Analysis System

Through a collaboration with IQVIA and Health Data Insight CIC, using data from the Cancer Analysis System (CAS) database in the United Kingdom, we sought to apply the framework established in Pilot 1.0 to confirm previously observed associations between rwOS and potential proxy endpoints (rwTTD/TTNT) in a nationally sourced, population-level, dataset. The CAS study followed the *Friends* RWE Pilot 1.0 protocol and compared the original findings from six US data sources to a UK Cancer Registry (CAS database) to compare RWD in aNSCLC. CAS is a cancer registry that includes more than 99% of all cancer patients in England and contains data on patient and tumor characteristics, treatments, hospitalizations, and mortality.

This study supported the findings from the original RWE Pilot 1.0 project and demonstrated a high level of correlation between rwOS and other rw endpoints, indicating the potential use of rwTTD/TTNT as a proxy endpoint for OS in real-world studies.

RWE Pilot 1.0: RWE Framework in Melanoma Patients from the RIC-Mel Database

In collaboration with Owkin/Centre Hospitalier Universitaire (CHU) de Nantes Pilot, we investigated the broader applicability of the RWE Pilot 1.0 framework in patients treated with immune checkpoint inhibitors for melanoma (anti-PD-1 monotherapy or anti-PD-1 combined with anti-CTLA-4 therapy as a first line of treatment or as a later line of treatment in advanced and metastatic melanomas). The project utilized the RIC-Mel database, which federates key patient information across 49 research institutions in France, with near comprehensive coverage and data that is highly curated and harmonized as all data collection is unified under a common CRF and digital platform interface for melanoma patients. Extending the RWE framework to patients with melanoma also provided an opportunity to align on data quality, standards, and investigate rw endpoints and their correlation to OS in other disease settings outside NSCLC. The applicability of the RWE framework in melanoma, supports further development of the framework as the structure for future studies.

RWE Pilot 2.0: Treatment Comparisons Analysis

Building on the work of the RWE Pilot 1.0, in 2019, *Friends* convened ten data partners, including organizations with data from EHRs or insurance claims, to conduct a parallel study where the different data sources were used to assess endpoints among aNSCLC patients receiving different first-line treatment regimens. Given the accumulating clinical experience with immune-oncology (IO) therapies, the RWE Pilot 2.0 was performed to assess treatment effect between platinum doublet chemotherapy, PD-L1 monotherapy, and PD-L1 in combination with platinum doublet chemotherapy using a common protocol. Patients meeting broad inclusion and exclusion criteria (treated with a qualifying therapy in first-line for aNSCLC, see SAP) were included to reflect the real-world population represented by the different data sources. Results of the study trended towards better outcomes in patients receiving IO over chemotherapy, directionally consistent with the findings of recent clinical trials.

RWE Pilot 2.0: Internal Consistency Analysis

Five RWE Pilot 2.0 data partners with data sourced from EHRs formed a subgroup to assess the consistency of findings in relation to trial results that examined the same treatment effects. Using the initial RWE Pilot 2.0 protocol as the basis for this study, additional patient inclusion/exclusion criteria were applied, leveraging EHR and lab data. The criteria, based on the KEYNOTE-189 clinical trial (platinum doublet chemotherapy versus PD-L1 in combination with platinum doublet chemotherapy in first-line aNSCLC), guided cohort selection to compare treatment effects in a more homogenous 'trial-like' real-world population, using rw endpoints of OS, TTNT, and TTD. The inclusion/exclusion criteria in this analysis were selected to facilitate greater alignment of baseline characteristics across datasets and greater similarity to clinical trial populations, although significant differences remain, on which approval of immune checkpoint inhibitors had been based (no balancing or weighting was applied). Treatment effects were compared at multiple restriction steps by select trial-based criteria, to assess how inclusion/exclusion criteria may have contributed to differences in observed treatment effect estimates. Additionally, the application of nuanced trial-based inclusion/

exclusion criteria to EHR data yielded important insights regarding data capture and the ability of RWD to identify precisely defined patient populations and characteristics.

The results from the above analyses were shared amongst all participating groups, to facilitate discussion of the combined learnings, and to subsequently develop a list of considerations for the design, conduct and interpretation of RWD studies from different data sources. Manuscripts are pending for each of the four expansion pilots.

Box 1 Pilot 2.0 Research Questions

- **Objective 1:** Describe demographic and clinical characteristics of patients with aNSCLC receiving frontline chemotherapy doublet, PD-(L)1 monotherapy, or PD-(L)1 + doublet chemotherapy.
- **Purpose:** Provide baseline understanding of the similarities/differences among the datasets to describe what confounding factors may need to be considered when interpreting the data.
- **Objective 2:** Evaluate treatment effect size in front-line therapy regimens using real-world endpoints.
- **Purpose:** Agree on data-source specific definitions and measurement of endpoints assessed through real-world data, in order to ensure reliability, consistency, and preservation of clinical meaning.

Considerations for a Real-world Evidence Framework

These five RWE pilot programs have yielded important lessons learned regarding establishing a RWE framework across multiple data partners in order to answer a common clinical question and we summarize these below.

Establishing a Research Question

To begin, defining and aligning on the clinical research question or objective is the key for any RWD study. All subsequent study considerations, including whether a data source is fit for purpose (meets certain data quality and completeness to address a specific question) and acknowledging potential limitations of data collection in real-world practice settings), will be guided by the clinical question. The considerations addressed in this whitepaper reflect lessons learned in the context of the RWE Pilot 2.0 research questions (Box 1).

Standardizing a common set of data elements

Establishing a core set of data elements to collect and standardize definitions could enable greater comparability across RWE studies, independent of data source(s). Demographics such as age and sex are minimal, structured, data elements that are typically readily available across independent data sources. However, eligibility criteria and definitions for other data elements demand thoughtful consideration and transparency such as: a) variables available in different formats (for example, PD-L1 biomarker positive/negative indicator vs. percent staining), b) variables requiring a curated definition (for example, ICD codes vs. lab values in the definition of organ function), or c) variables requiring extraction from unstructured data (for example, status of advanced cancer at initial diagnosis vs. progression after initial, earlier-stage diagnosis).

Using the RWE Pilot 2.0 as a case study, we propose a core set of data elements for consideration in real-world oncology studies **(Appendix)**. We include further considerations for harmonizing definitions across datasets with the prerequisite core data elements to address the pre-specified research question.

Considerations

1. Identify a core set of data elements that can be systematically defined across real-world data sets for the proposed study.

Important considerations when creating a core set of data elements include commonality and availability of data across datasets, the clinical setting, and study objectives. Data elements that are consistently available across all or the majority of datasets will reduce data variability and increase understanding of data missingness within the datasets. The completeness of each data element within each dataset, as well as across datasets, should be evaluated and reported. Selection of core data elements should also consider the clinical context. As a result, a core set of data elements **(Appendix)**, will require modifications when applying across disease or therapeutic class. For example, smoking status may be relevant across multiple diseases but provide particularly important prognostic information in lung cancer, as opposed to other cancer types, such as melanoma. The phenotype of melanoma requires different information such LDH, BRAF and histologic details of primary lesion. However, age, sex, and stage of disease are important data elements across all of the RWE Pilots.

Additionally, consider that the prognostic value of a characteristic such as smoking status will depend upon the level of variable completeness and definition used for this data element (for example, patient was never a smoker vs. there is no evidence of smoking history). Last, consider the study objectives and endpoints to be measured when selecting core data elements as this will help with selecting the most appropriate data elements. For example, patient age at advanced diagnosis would be a more appropriate characteristic than age at initial diagnosis (where a patient presented with early stage disease and now has aNSCLC) for a study objective to observe treatment effect in patients with aNSCLC.

2. Identify the analytic variables that require a high level of harmonization vs. those that can accommodate variability across data sources.

Harmonized definitions should be employed wherever possible and particularly for data elements with high likelihood of impact on endpoint calculations. However, standardized definitions are not always feasible and variability across datasets may be acceptable as assessed on a per study basis. For example, even when using a common case report form across institutions within a data source, heterogeneity of information reporting can persist between institutions (e.g. lymph node removals can be coded differently depending upon the clinical site). To identify data elements where variability may be acceptable, consider 1) whether harmonization is possible given each source of the data element (e.g., EHR vs. claims data; diagnosis vs. laboratory value for defining a comorbidity) and the underlying population and 2) whether harmonization is necessary. For example, treatment initiation date could be sourced from administrative claims, an electronic prescription order, or date of administration within an EHR, and a flexible definition could ensure more comprehensive identification of patients receiving a particular treatment. Similarly, practice patterns can vary across geographic regions and clinical practices and can impact how a frontline therapy is defined within different datasets. Flexible definitions may be needed that place greater emphasis on accurate identification of appropriate patient populations within the context of each dataset compared to harmonization of variable definitions among datasets. Last, a harmonized definition may not be necessary for some data elements, particularly where little to no impact on the included patient population or calculation of endpoints is expected or where stringent definitions could limit potential observations. For example, the RWE Framework broadly identified inclusion based upon treatment with a platinum doublet chemotherapy or IO monotherapy or IO in combination with any platinum doublet chemotherapy but did not restrict to specific drugs or pre-defined procedure codes (Healthcare Common Procedure Coding System (HCPCS)/Drug Codes) for specific regimens. As a result of allowing for inclusion based upon a class of drugs, the study evaluated on and off-label use that might have been excluded from the study if more stringent inclusion criteria had been applied.

Similarly, if the variable in question is included in an analysis as a potential confounder (i.e., to adjust for confounding), the specific form the confounder takes in the model may be less relevant than other variables for which specific inferences are intended. The strength of confounding exhibited by each variable is also important and consistent modeling of each variable across datasets will be important to control for confounding.

3. Align on harmonized definitions where appropriate.

Harmonizing the definition for key data elements can help account for variability likely to exist among datasets in terms of the source and patient population represented. Factors to address when aligning definitions include accuracy, extent of missingness, and granularity. Similarly, categories of reference values for classification of covariates (such as lab values) should be agreed upon and used consistently.

First, data accuracy is important to consider for harmonization of definitions. For example, the definition of covariates such as organ function, which can be extracted from ICD codes or laboratory test results, should be considered for potential implications on results. Analyses utilizing extracted laboratory values are likely to have greater granularity when comparing magnitudes of organ failure (normal, mild, moderate, or severe) as compared to definitions based on structured ICD codes, which may communicate less information (for example, evidence of organ disease) but be recorded more frequently than lab test results. Different considerations for missingness should be accounted for when comparing diagnostic data [ICD codes] vs. lab value data, where absence of ICD codes or test results does not equate to absence of a condition. Differentiation between patients with no evidence of organ disease and patients with unknown organ function will be difficult or impossible, particularly when utilizing only structured ICD codes.

Second, consider the source/level of detail of data elements when harmonizing definitions, particularly where there is variability in how the data element is documented. For example, PD-L1 expression can be reported in a variety of ways (pathology report vs physician reported) and additionally have different thresholds for what constitutes a “positive” or “negative” result. When both sources of biomarker status are used, definitions should reflect the existing variation and attempt to align populations where possible.

Third, when addressing data missingness and the reason for missingness (whether or not missingness is at random) it is important to understand the indication(s) for measuring covariates such as organ function. While certain tests of organ function may be done routinely in line with clinical guidelines, some may be ordered specifically if patients have preexisting conditions or present with certain symptoms. In that case, the ascertainment is biased and will impact endpoint estimates. Similarly, HIV testing is not routinely done outside of clinical trial selection. It is also important to consider that for some comorbidities, such as hypertension, using diagnostic vs. lab data for identification could lead to different endpoint estimates. Patients who have hypertension that is controlled through medication may be identified by ICD codes for hypertension or diabetes in some patient records. However, those same patient records would indicate normal blood pressure values due to control with anti-hypertensive medications. As a result, the use of diagnostic codes vs. lab values may not yield the same value for some covariates. The same would apply to use of diagnostic codes vs lab values of blood glucose to identify diabetes in a patient on medications to normalize blood glucose level. The ascertainment window for laboratory values will also impact this measure, taking into account proximity of data ascertainment to timeframe of interest and how to address reporting of multiple laboratory values during the study period.

Last, consider the granularity of definitions. For example, identification of adverse events is of particular importance in RWE but is especially difficult to measure if relying primarily on structured data to attribute to a particular therapy or treatment. Assignment of attribution requires chart review, which is variable and time consuming. A related example is use of the term “advanced” (this includes both Stage IIIB/C and IV disease) to identify aNSCLC patients, which can have an impact on observed endpoints for the specified population. The term “advanced” may be defined as a patient with a certain stage of disease at diagnosis or having developed metastatic disease independent of initial stage, but if the focus of the study is the treatment of metastatic NSCLC then only Stage IIIB/C patients should be eligible if they progressed with metastatic disease. The definition of advanced in this case will depend upon the ability of each dataset to capture and identify progression as a disease indicator or as a clinical endpoint within each dataset. Similarly, large amounts of missingness in progression data may also impact the patient population selected for a study and should be considered when aligning definitions and interpreting results. Other considerations include clinical guidelines and workflows for disease surveillance (following treatment), that may or may not differ across practice settings, and duration of follow-up.

4. Review of the distribution of identified variables by collaborators.

Lastly, review the distribution of the pre-specified variables for each population across datasets as an internal check on the study definitions and alignment on methods implementation. Specifically, use this informal assessment to identify unexplained outliers associated with a collaborator/data source (e.g., high number of early deaths or very long survival, or high percentage of advanced diagnosis) and missingness within datasets. This can help to not only identify where additional checks are needed to determine if an error has been made (e.g., in the harmonization process) but data sources to exclude for certain analyses because data is not fit-for-purpose (at least for the study at hand).

Methodological considerations for interpreting endpoints

The information that can be gleaned from a real-world study depends on the methodology and definitions used to select the patient population. In addition to a core set of data elements and harmonized definitions for patient selection, it is essential to align on common statistical methodology for analyzing and interpreting endpoints. The level of specification in the real-world methodology is paramount as this will help to reduce confounding and variability in implementation due to differences in interpretation of the protocol. Ultimately, a thorough methodology will be important to facilitate earlier engagement with regulatory agencies for RWD to support drug development and regulatory decision making as well.

- **Identify and summarize the source of the endpoint information and how the endpoint was derived.**

A thorough understanding of the source of the data used to derive the endpoint, including limitations and completeness of the data, is necessary to draw accurate conclusions. The source, completeness, and accuracy of mortality data in observational studies can impact comparative effectiveness inference. For example, death information is not systematically captured in routine clinical care in the U.S., potentially requiring multiple sources of information to be used to capture mortality. Further, calculating sensitivity and specificity for mortality in EHR data requires linking to a gold standard and may present challenges for de-identified data sources, particularly in the U.S. data sources. Even though a centralized mortality database, the National Death Index, exists for research use, it is often not linked to in the context of RWD, and regardless is not complete in a timely manner, precluding its use for evaluation of new therapies. Reporting metrics including data completeness, sensitivity, and specificity for mortality should be established. In cases where EHR data is used, more information can be obtained by performing chart review as opposed to strictly depending on structured data.

- **Determine appropriate endpoints.**

The specific research question and clinical context will drive selection of an appropriate endpoint **(Tables 1 and 2)**. For example, objective endpoints, such as OS are susceptible to factors such as post baseline events, such as treatment crossover, and may make treatment effects harder to interpret. OS may also suffer from substantial missingness. Some endpoints such as progression free survival or overall response rate may not be appropriate for RWE studies due to the difficulty of identifying progression or response in structured RWD. Specifically, progression/response is not consistently reported in real-world care and, where it is available, requires time consuming chart review to extract. Further, while clinical trials rely upon objective and well-defined Response Evaluation Criteria in Solid Tumors (RECIST) criteria to measure progression, in RWD, progression/response assessments may not be as rigorous as RECIST, with more subjective clinician interpretation, variability in the scheduling of imaging tests than in trials, and less rigorous reporting in the EHR. Consensus in how to define and document progression/response in structured data would make this endpoint more readily available and appropriate as a rw endpoint.

Conversely, treatment-based endpoints, such as rw time-to-treatment discontinuation (rwTTD) or time-to-next treatment (rwTTNT), are more objective, may be more readily interpreted, and may present advantages regarding completeness. However, these endpoints do not explicitly capture differences in drug effectiveness. Interpretations are complicated by the diversity of reasons for treatment discontinuation or switch (such as toxicity or patient preference), as well as differences in expected treatment duration (e.g. pre-defined number of cycles vs indefinitely) across therapies or indications. For example, treatment is arbitrarily stopped after 4-6 cycles in some cancers regardless of the status of disease. Similarly, it is difficult to assess the end date for oral oncolytics using only structured data from EHRs. Despite different reasons for discontinuation, the clinical endpoints may be more relevant to the patient. The research question, clinical context, quality of mortality variable, and availability of additional data (e.g., on post-baseline therapies or reasons for treatment discontinuation) should help guide endpoint selection.

- **Provide transparency on endpoint derivation, definition and transformation.**

Transparency regarding how endpoints are derived (e.g. detailed documentation of deviation in methodology regarding the source of the data and what transformation is conducted to derive the endpoint) is important for 1) standardizing methodology and confirming comparability of results, 2) performing validation studies of the endpoints, and 3) building trust in the results of RWD studies. Variability in data sources, completeness and quality of data, as well as limitation of analysis plans, including defining exposure, endpoints, and key covariates, and potential resulting biases all need to be considered.

It may be preferable, when comparing data from disparate RWD, to pre-specify more than one estimate or measure of association for comparison (for example, proportion of patients that are event-free at pre-specified timepoints [the survival function] and a hazard ratio) **(Table 2)**. Various measures may be affected differently by dataset characteristics and study design elements, including distribution of exposure to treatments over time, duration of follow-up per treatment arm, and crossover from one treatment arm to another. Characterization of adjusted survival curves can be considered.

- **Ensure comparability of index dates**

When conducting real-world studies to assess treatment effect, the comparability of index date (e.g., the start of the time when patients experience the qualifying event and become at-risk of having the endpoint of interest) is of critical importance (Box 2). For example, if entry into a dataset is linked to post-baseline data, this may generate bias and render estimates not comparable across data sources.

- **Ensure comparability of censoring rules and event dates**

Harmonization, and transparency where harmonization is not possible, of censoring rules and event dates will increase interpretability of results. Important considerations include the length of follow-up available (and continuity of follow-up), the therapy being investigated, and the endpoints to be measured **(Table 1)**. Various types of patient activity recorded in different datasets (e.g., structured visits, labs, abstracted dates) and their applicability for use as censoring dates for the chosen endpoint, should be considered. For example, structured visit or claim dates might be most readily available and could be standardized for a mortality endpoint, but endpoints such as progression-free survival require a finer distinction between types of clinical encounters.

A data cutoff should be pre-specified, to ensure ascertainment of event and censor dates over a standard timeframe. The selection of data cut-off should be informed by the research question (how much follow-up is expected to be necessary to observe a treatment effect for the disease?) and the endpoint selected (how much follow-up time is necessary to accrue a meaningful number of events, and, in case of mortality, for datasets to optimize sensitivity of event capture from external data sources?).

When defining an event where a combination therapy is used, censoring rules must be harmonized across datasets to ensure consistent assignment of discontinuation (e.g., do both therapies within a combination need to be discontinued or does discontinuation of a single therapy within a combination constitute an event?).

Box 2
RWE Pilot Definitions for Identification of Index Date

- Frontline will be defined as the first systemic drug regimen given to NSCLC patients subsequent to the patient's date of advanced diagnosis and after the start of time period of the study.
- Frontline will begin on the date that the systemic cancer treatment was initiated and include all agents received within 30 days following the day of first administration.
- The index date will be based on the earliest drug episode (e.g., first administration or non-cancelled order) of the frontline therapy for advanced disease.

- **Assess "fitness-for-purpose" to increase confidence in endpoint.**

Studies to assess the reliability of the endpoint used are necessary until a larger body of RWE comparative effectiveness in oncology literature is available. These studies may confirm accuracy of OS (i.e., identify if differences in OS estimates are true or artifactual due to incomplete mortality data) or support correlation of a proxy endpoint to a gold-standard endpoint.

- **Contextualize methods and results against other data sources, as applicable.**

It is important to consider factors that increase confidence in the real-world measure/endpoint in RWD. This could also be addressed by examining associations of each covariate (e.g., age, sex, etc.) with survival endpoints to assess whether observed effects fall in line with expectation and whether their directions and magnitudes of association are comparable across study populations. Further, where possible, comparability of survival curves for the selected endpoint to similar epidemiological studies performed within similar populations, such as with SEER data, should be conducted. However, this requires confidence that the study population is similar to the patient population in the comparison population. This may be difficult to achieve given the possible number of patient characteristics that are un-identifiable or confounding within an observational dataset. Potential for confounding also contributes to the lack of agreement or inability to conduct appropriate RWD-based analyses of clinical trial results, and in addition to a lack of endpoint comparisons as described above, as PFS assessment is often not readily available in RWE. Similarly, clinical trials involve detailed protocols for care delivery (e.g., how standard of care or the investigational agent is delivered) whereas differences in real-world protocols can confound comparability among studies. Where possible, additional inclusion/exclusion criteria, as well as weighting, can be applied to better align real-world data populations in comparison to other data sources, (such as clinical trial or other observational data) to enhance comparability of findings. Ongoing evaluation of the study objectives and endpoints, and revisions, where necessary, should occur throughout the study to promote comparability.

- **Ensure replicability of endpoint measures.**

Similar to comparability to other studies, efforts should be taken to increase replicability of endpoint measures. For example, endpoint measures such as treatment administration date or date of service to derive a time to next treatment may be readily identified in RWD and, thus, can be replicated across datasets as compared to identification of progression, which may require date of and results of radiographs, laboratory tests, and/or clinician assessments. It is essential to conduct sensitivity analyses when comparing across datasets because of potential variability to ensure robustness of the results and stability of the estimates, especially with variance in statistical methodologies which may account for differences.

A process for assembling “fit-for-purpose” real-world datasets

Although there exists certain challenges and limitations with RWE, with a thorough understanding of the data provenance and well-designed study protocols, real-world datasets can be assembled that produce robust analyses that complement those of clinical trials and other datasets. By applying the *Friends* RWE Framework in several clinical settings and diverse data sources, we developed a process for assembling a “fit-for-purpose” real-world dataset to guide future real-world studies **(Figure 1)**. It is important to emphasize that the recommendations enumerated in this whitepaper are intended to inform a process for assembling fit-for-purpose datasets and methodologies based upon lessons learned from the *Friends* RWE collaborations. The exact core data elements, definitions, and protocols may not necessarily apply to all clinical settings or datasets as the RWE Pilots were developed to inform treatment effect of IO therapies in a specific disease setting. Certainly, modification of this framework will be appropriate to expand to other drug classes, other cancer types, different health systems (international studies), and beyond oncology. Specifically, consideration of the added complexity is necessary to adapt a standard protocol across multiple settings, e.g., different study periods due to different scope/timeline of regulatory approvals and existence of different regulatory and clinical guideline bodies, as well as considerations around accessing sensitive patient data under different patient privacy restrictions or using de-identified patient data. With more widespread application of this framework, we can begin to accumulate a body of evidence in support of various real-world endpoints and inform regulatory policy.

Table 1: Real-world Endpoint Definitions

rwEndpoint	Definition	Censor Date	Considerations for Definitions and Alignment
rwOS	Length of time from the index date to the date of death, or disenrollment (need to define gap in enrollment). For claims data, health plan disenrollment date is incorporated if deaths are not captured among those who leave health plan coverage.	Last structured recorded clinical activity within the real-world database including prescription, office or institutional billing claims data, or end of follow-up period, whichever occurs earliest.	• Definition variability appropriate. • Separate definitions required for EHR-based vs claims-based data sources for all endpoints. • Consider the completeness of vital status data.
rwTTNT	Length of time from the index date to the date the patient received an administration of their next systemic treatment regimen or to their date of death if there is a death prior to having another systemic treatment regimen.	Last known activity or end of follow-up.	• Length of patient follow-up to capture subsequent treatment regimens.
rwTTD	Length of time from the index date to the date the patient discontinues frontline treatment (i.e., the last administration or non-cancelled order of a drug contained within the same frontline regimen). Discontinuation is defined as having a: • having a subsequent systemic therapy regimen after the frontline treatment; • having a gap of more than 120 days with no systemic therapy following the last administration; or, • or having a date of death while on the frontline regimen.	Last known usage (i.e., administration or non-cancelled order) of frontline treatment.	• Consider standard duration of frontline treatment regimen.

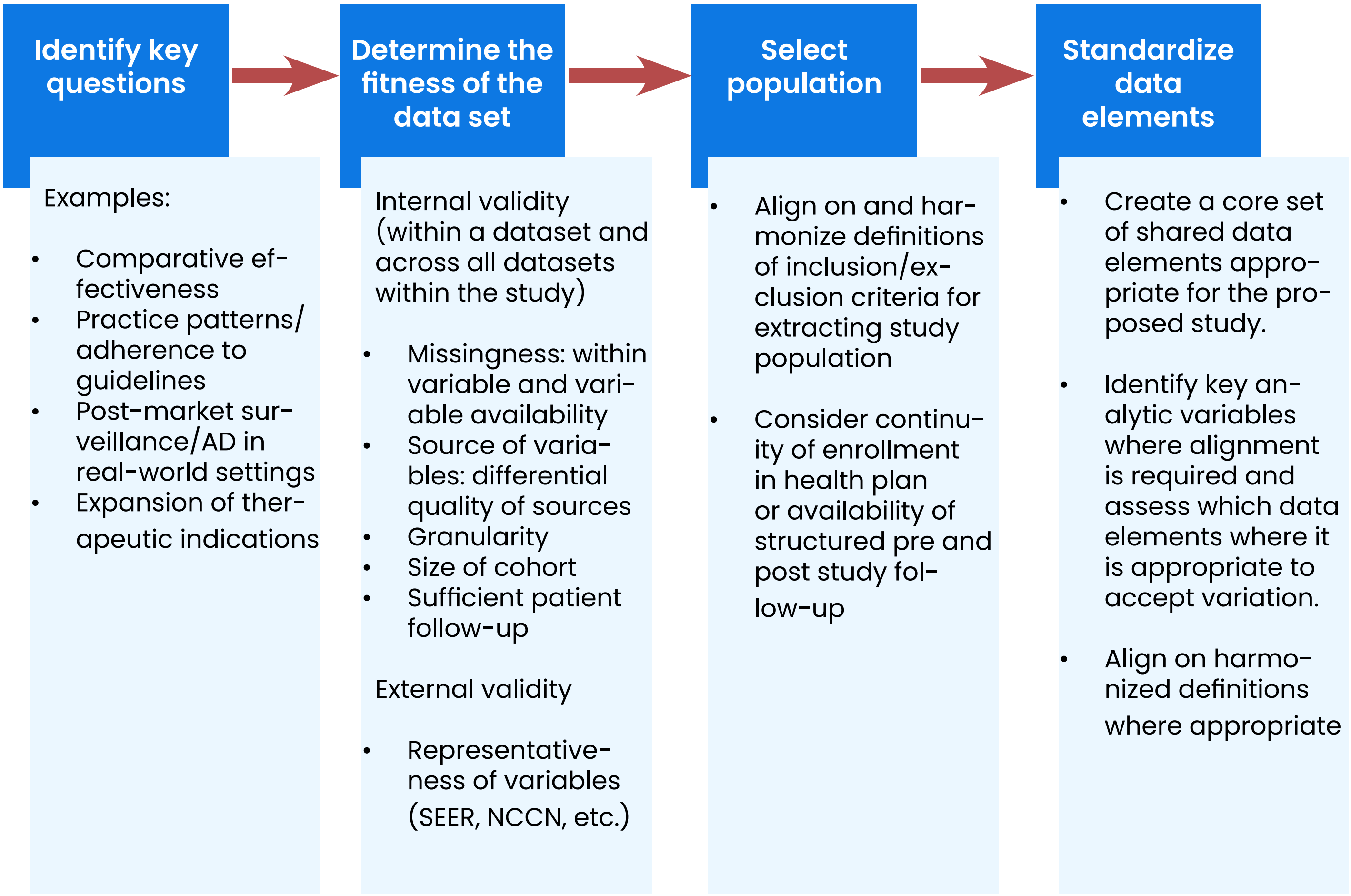
Table 2: Comparison of rwEndpoint Event Estimates to Assess Treatment Effect

		rwOS				rwTTD				rwTTNT			
Group	Treatment Regimen	N of Patients	12 month Survival Estimate	95% Confidence Interval (lower)	95% Confidence Interval (upper)	N of Patients	12 month Treatment Continuation Estimate	95% Confidence Interval (lower)	95% Confidence Interval (upper)	N of Patients	12 month Next Treatment Estimate	95% Confidence Interval (lower)	95% Confidence Interval (upper)
A	Platinum Doublet Chemotherapy	1542	0.530	0.504	0.555	1542	0.035	0.026	0.045	1542	0.230	0.208	0.252
	PD-(L)1 Monotherapy	257	0.501	0.428	0.569	257	0.322	0.253	0.393	257	0.409	0.340	0.476
	PD-(L)1 + doublet chemotherapy	132	0.506	0.401	0.602	132	0.330	0.229	0.439	132	0.435	0.333	0.532
B	Platinum Doublet Chemotherapy	697	0.556	0.518	0.592	1572	0.047	0.033	0.065	1572	0.246	0.215	0.279
	PD-(L)1 Monotherapy	429	0.615	0.567	0.659	492	0.263	0.223	0.306	492	0.440	0.392	0.486
	PD-(L)1 + doublet chemotherapy	233	0.570	0.503	0.630	233	0.227	0.176	0.283	233	0.437	0.373	0.500
C	Platinum Doublet Chemotherapy	997	0.629	0.596	0.661	998	0.042	0.030	0.056	995	0.216	0.189	0.244
	PD-(L)1 Monotherapy	160	0.575	0.470	0.667	160	0.008	0.001	0.039	159	0.286	0.206	0.372
	PD-(L)1 + doublet chemotherapy	56	0.734	0.565	0.846	56	--	--	--	56	--	--	--
D	Platinum Doublet Chemotherapy	1351	0.494	0.467	0.521	1351	0.033	0.025	0.044	1351	0.230	0.207	0.253
	PD-(L)1 Monotherapy	235	0.515	0.448	0.577	235	0.191	0.144	0.244	235	0.315	0.256	0.375
	PD-(L)1 + doublet chemotherapy	75	0.456	0.339	0.565	75	0.187	0.108	0.282	75	0.290	0.190	0.396

E	Platinum Doublet Chemotherapy	1146	0.461	0.432	0.489	1146	0.044	0.033	0.057	1146	0.180	0.159	0.203
	PD-(L)1 Monotherapy	90	0.542	0.433	0.638	90	0.216	0.136	0.307	90	0.340	0.242	0.439
	PD-(L)1 + doublet chemotherapy	33	0.636	0.450	0.775	33	0.394	0.231	0.554	33	0.455	0.282	0.612
F	Platinum Doublet Chemotherapy	9707	.430	.420	.440	9707	.213	.204	.222	9707	.185	.177	.193
	PD-(L)1 Monotherapy	--	--	--	--	--	--	--	--	--	--	--	--
	PD-(L)1 + doublet chemotherapy	--	--	--	--	--	--	--	--	--	--	--	--
G	Platinum Doublet Chemotherapy	1453	0.582	0.553	0.612	1453	0.039	0.029	0.053	1453	0.231	0.208	0.257
	PD-(L)1 Monotherapy	132	0.657	0.565	0.763	132	0.253	0.164	0.393	132	0.461	0.367	0.580
	PD-(L)1 + doublet chemotherapy	57	0.700	0.577	0.848	57	--	--	--	57	0.507	0.381	0.675
H	Platinum Doublet Chemotherapy	4752	0.463	0.447	0.478	4752	0.039	0.033	0.046	4752	0.170	0.159	0.182
	PD-(L)1 Monotherapy	601	0.594	0.550	0.636	601	0.319	0.278	0.361	601	0.442	0.398	0.485
	PD-(L)1 + doublet chemotherapy	243	0.564	0.495	0.628	243	0.319	0.256	0.382	243	0.410	0.343	0.475

*-- data unavailable due to small sample size, insufficient follow-up, or lag in availability of data within data source

Figure 1. Step-wise approach to assembling "fit-for-purpose" real-world dataset.



Identify key questions

Examples:

- Comparative effectiveness
- Practice patterns/ adherence to guidelines
- Post-market surveillance/AD in real-world settings
- Expansion of therapeutic indications

Determine the fitness of the data set

Internal validity (within a dataset and across all datasets within the study)

- Missingness: within variable and variable availability
- Source of variables: differential quality of sources
- Granularity
- Size of cohort
- Sufficient patient follow-up

External validity

- Representativeness of variables (SEER, NCCN, etc.)

Select population

- Align on and harmonize definitions of inclusion/exclusion criteria for extracting study population
- Consider continuity of enrollment in health plan or availability of structured pre and post study follow-up

Standardize data elements

- Create a core set of shared data elements appropriate for the proposed study.
- Identify key analytic variables where alignment is required and assess which data elements where it is appropriate to accept variation.
- Align on harmonized definitions where appropriate

Appendix: Core Data Elements

Variable Name	RWE Protocol Definition	Considerations for Definitions and Alignment
Minimal Structured Data Elements		
Characterization of the study population		
Advanced diagnosis date	● Date of diagnosis of advanced disease: ○ Initial diagnosis with stage IIIB, IIIC, or IV or ○ First recurrence/progression after earlier stage diagnosis	● Definition should be aligned or transparency of variability must be provided ● Dependent upon ability to identify recurrence or progression within dataset ● Protocol for identification of progression may differ across datasets
Age at index	● Age at the start of frontline therapy, as previously noted in the index date definition ● Reported as continuous, categorical, and binary. ● Categorical: ○ <49 years ○ 50-64 years ○ 65-74 years ○ 75+ years ● Binary: ○ <75 years ○ 75+ years	● Binary categorization should reflect study objectives. If comparison to clinical trial data is of interest, binary age should be reflective.
Sex	● Male ● Female ● Other / Unknown	
Region	● Based on patient's state of residence: ○ Midwest ○ Northeast ○ South ○ West ○ Other/Missing	

Race	● White ● Black or African American ● Asian or Pacific Islander ● Other (Native American, Alaskan Native) ● Unknown/Missing	● The level of harmonization for this covariate should consider key objectives for the study as granularity could impact endpoints.
Histology	● Non-squamous cell carcinoma ● Squamous cell carcinoma ● NSCLC histology not otherwise specified (NOS)	● The level of harmonization for this covariate should consider key objectives for the study as granularity could impact endpoints.
First-line regimen	● First regimen patient received in the advanced setting. Eligible frontline therapies include: ○ Platinum Doublet chemotherapy (cisplatin, carboplatin, oxaliplatin, or nedaplatin with pemetrexed, paclitaxel, nab-paclitaxel, gemcitabine) ○ PD-(L)1 monotherapy (pembrolizumab, nivolumab, atezolizumab) ○ Any PD-(L)1 + doublet chemotherapy combination (pembrolizumab, pemetrexed and platinum or pembrolizumab, platinum and paclitaxel or nab-pacltiaxel)	● Variability in defining a frontline regimen may be acceptable depending upon clinical practice variation that may exist, particularly if using internationally sourced data, while minimizing misclassification of exposure. ● Consider whether granularity regarding the exact drug or class of drug is important. ● Consider how distinguish monotherapy vs combination therapy
Clinical characterization of the study population		
Smoking status	● Patient's smoking status as documented at any point prior to the data cutoff for this study: ○ History of smoking ○ No history of smoking ○ Unknown/not documented	● A more granular definition such as smoking status at diagnosis or at frontline treatment initiation (index date) may be a better barometer. ● "History of smoking" does not distinguish between current vs. former smoking status.

Group stage	● Stage of disease at the time of initial diagnosis with NSCLC: ○ 0 ○ I ○ II ○ III - IIIa - IIIb - IIIc ○ IV ○ Group stage is not reported	
PD-L1 tested (before or 30 days after the index date)	● Tested for PD-L1 ○ Tested ○ Untested ● **NOTE:** Testing may occur at any point before or up to the index date, defined previously. ○ Where available: the test date will be identified as the most recent date available across the “specimen collected” date, “specimen received” date, and “result date” variables.	
PD-L1 status (before or 30 days after the index date)	● Result for PD-L1 test among those with documented testing ○ PD-L1 positive ○ PD-L1 negative/not detected ○ PD-L1 equivocal ○ No interpretation given in report ○ Results pending/Unknown ● Testing may occur at any point before or up to the index date.	● Alignment necessary for testing time-frame ● Consider study objectives when defining look-back and cut-off dates. ● This result is based on test interpretation as reported to the physician, which may consist of differing cut-offs.
PD-L1 staining (before or 30 days after the index date)	● Staining level result for PD-L1 test among those with documented testing ○ <1% ○ 1% - <50% ○ ≥50% ○ Unknown ● **NOTE:** Testing may occur at any point before or up to the index date, defined previously.	● Alignment necessary for testing time-frame ● Consider study objectives when defining look-back and cut-off dates. ● Unstructured data may not be widely available. ● Greater precision is available but extent missingness should be addressed.

Organ function at index	● Patient's renal/hepatic function at the index date, as previously described, based on structured lab data ○ Severe renal/hepatic failure ○ Moderate renal/hepatic failure ○ Normal renal/hepatic function ● **NOTE:** Restricted to patients with creatinine serum lab values for renal function or total bilirubin, aspartate aminotransferase (AST), or alanine transaminase (ALT) for hepatic function up to 30 days before the index date ○ Categorization of renal function defined as: ○ Severe: >3x upper limit of normal ○ Moderate: 1.5-3x upper limit of normal ○ Mild: >ULN-1.5x ULN ○ Normal: ≤ULN ○ Categorization of hepatic function defined as: - Severe defined by one of the following: ● Total bilirubin >3x upper limit of normal ● AST >5x upper limit of normal ● ALT >5x upper limit of normal - Moderate defined by one of the following: ● Total bilirubin 1.5-3x upper limit of normal ● AST 3-5x upper limit of normal ● ALT 3-5x upper limit of normal - Mild defined by one of the following: ● Total bilirubin >ULN-1.5x ULN ● AST >ULN-3x ULN ● ALT >ULN-3x ULN - Normal defined by meeting none of the above criteria	● Alignment necessary for testing time-frame and definition of severity categories. ● Consider study objectives when defining look-back and cut-off dates. ● If data from more than one lab test are available for a given individual, use the most recent value.

Presence/ab-sence of chronic organ disease (ICD9 code)	● Defined as at least one diagnostic code any time prior to and including the index date, defined as having one of the following ICD9/10 codes: [N18.x, N19.x.] or [585.X, 586.X] for kidney disease or [K70-K77] or [570.X, 571.X, 572.X, 573.X] for liver disease. ○ Yes ○ Unknown	● Alignment of ICD codes required ● Limited in ability to distinguish no organ disease vs no evidence of organ disease. ● See SAP for complete list of ICD codes used.
Performance status (ECOG) at index	● Patient's ECOG status at the time of the index date, defined previously ○ 0 ○ 1 ○ 2+ ○ Unknown ● **NOTE:** ECOG may have been recorded up to 30 days prior to the index date, OR up to 7 days after the index date, whichever is closest to the index date. If there are multiple ECOG values at the same absolute distance from the index date, priority is given to the ECOG value that precedes the index date. For patients with multiple ECOG values recorded on the same day, the highest value will be selected.	● Alignment necessary for time-frame ● Consider study objectives when defining look-back and cut-off dates.
CNS Metastases	● Defined as at least one diagnostic code up to 30 days after the index date. ● Complete list of ICD codes was created. ● Secondary CNS neoplasm codes: ○ ICD9 codes: 198.3, 198.4/ ○ ICD10 codes: C79.31, C79.32, C79.49. ● Primary malignant neoplasm codes should be included if they occur after index because the likelihood of a CNS primary after a metastatic NSCLC is very unlikely and miscoding is common. Primary CNS neoplasm codes: ICD9: 191.-191.9/ ICD10: C71.0-C71.9	● Alignment necessary ● Consider structured ICD codes. ● See SAP for complete list of ICD codes used.
Endpoints		
Date of death	● rwOS event date	● Alignment necessary ● Consider minimum standards for data completeness
Start date of regimen after frontline	● Start date of regimen immediately after frontline (i.e., second-line) ● **NOTE:** Patients with a death prior to having another line will be considered as having an event ● rwTTNT event date for endpoints of regimen after frontline [i.e., second-line] or death	

Last confirmed activity date	● Patient’s last known structured recorded clinical activity ● For calculation of structured follow-up time ● rwTTNT or rwOS censor date	● Definition variability appropriate. Separate definitions required for EHR-based vs claims-based data sources. ● Various types of patient activity recorded in different datasets (e.g. structured visits, labs, abstracted dates) and their applicability for use as censoring dates for the chosen endpoint, should be considered
Frontline discontinuation date	● Date of frontline regimen discontinuation ● rwTTD event date	
Frontline last continuing date	● Date of last known frontline regimen when there is no frontline discontinuation (i.e., still on frontline therapy) at the data cutoff ● Or Last observed administration date of frontline. ● rwTTD censor date	
Additional Data Elements for Consideration		
Age at advanced diagnosis	● Age at advanced diagnosis (continuous)	
Ethnicity	● Hispanic ● Non-Hispanic ● Unknown/Missing	● Same considerations as for race
Median income (quartile)	● Median household income (zip-level quartiles) ○ 1 (lowest median household income) ○ 2 ○ 3 ○ 4 (highest median household income) ○ Unknown	
Other Biomarkers	● ALK/EGFR	● Consider clinical implications of additional biomarkers as exclusion criteria