Evaluating Whether RWD Is Suitable for Planning Eligibility Criteria and Supporting Recruitment

The questions below are intended to help sponsors determine whether available real-world data (RWD) sources can be used to support design of eligibility criteria and/or recruitment. These questions can be used to assess in-house data, as well as data from third-party providers. When assessing RWD for recruitment purposes, use these questions in conjunction with Evaluating Feasibility of RWD-Supported Recruitment.

Questions	Factors to Consider
Are important eligibility criteria identifiable in the data?	▶ **Which eligibility criteria can be identified from RWD.** Consider whether eligibility criteria that are important to the success of the study can be identified from structured data, as well as whether reliable queries can be run on unstructured data (e.g., via natural language processing). In many cases, it may be necessary to define proxy measures that best match planned eligibility criteria against the available data. Certain types of eligibility criteria will not be identifiable from electronic health records (EHR) or claims data at all. ▶ **Feasibility of designing appropriate queries.** In part because RWD sources often do not directly align with trial eligibility criteria, what seems like a relatively straightforward diagnosis may require querying against multiple potential RWD criteria. Developing these queries requires collaborative iteration and validation between individuals with clinical expertise and individuals with epidemiological expertise.
Are data of sufficient relevance and quality?	▶ **Acceptability of errors in data.** It is important to understand which errors may render the data less useful, and which can be managed. As an example, undercoding (i.e., not coding a condition that a patient has) may be more common than overcoding in EHR and claims data. Thus it is possible that more patients will meet the inclusion criteria (which is good for trial feasibility), *and* that more patients will meet the exclusion criteria (which is problematic for trial feasibility), than the RWD suggest. ▶ **Recency of data relative to study needs.** Claims data, for example, are often 30-90 days old; this will be adequate for some studies (e.g., identifying patients with chronic illnesses) and inadequate for others (e.g., finding newly diagnosed patients initiating a first treatment). Additionally, if combining RWD sources, it is important to consider implications of different data availability timeframes. Timing is often less critical for tasks such as estimating sample size availability or event rates than for patient recruitment. ▶ **Generalizability of data.** Many databases are limited to certain populations or geographic regions, and it is important to understand and account for these limitations when interpreting data for study planning purposes. For example, for a condition that is mostly found in a Medicare population, a commercial claims database may underestimate the prevalence of the condition. For multi-regional clinical trials, it may not be possible to identify a database with geographic coverage that fully matches the planned trial locations, and it is important to understand the extent to which data from one geographic region can be generalized to understand broader trends across regions.

Will data analysis be timely and cost-effective?

- **Analysis becomes more challenging as the number of databases increases.** The most efficient process is to run analyses on a single, centralized database. Analysis even on a site-by-site basis can still have value, but will take substantially greater time and resources to conduct. Similarly, when a centralized database does not already exist, it is important to consider the substantial effort required to develop one; common data models (e.g., PCORnet) that harmonize and standardize data across systems in advance can mitigate such challenges but may not be adopted by all relevant databases.
- **Interoperability challenges exist when pooling data.** Variability exists in the record systems used, as well as the coding used to reference various health events, who enters the data into the systems, the timing of data entry, and the meaning of some data elements. Data elements that appear similar across different RWD sources may not be the same, due to cultural and technical variability. Similarly, data elements that appear different across RWD sources may actually measure the same variables. Involvement of individuals familiar with the data source (e.g., individuals from provider organizations) can help to identify and manage such challenges.