Paper RE02

Supporting Clinical Trial Recruitment using Real-World Data

Yi-Chien Lee, Pfizer Inc., New York, NY, USA

ABSTRACT

Clinical trial recruitment is a resource-intensive process and can be even more challenging if no suitable subjects are identified after the site opens. Assessing feasibility before the study starts is crucial as it helps the project team design appropriate inclusion and exclusion criteria and plan the project timeline based on the estimated recruitment speed. In this poster, we use a recent project to showcase how to utilize real-world data to understand the target population in the US. Age and gender information is used to understand the target demographic. ICD-9/10 diagnosis codes are used to evaluate the medical history; NDC/HCPCS claims are used to capture treatment use; and LOINC code/test names/results are used to understand the patient's lab findings. By exploring the potential subject pool size using various inclusion and exclusion criteria, researchers can identify the optimal balance—targeting the appropriate population without being overly restrictive.

INTRODUCTION

Traditionally, feasibility questionnaires are distributed to randomized clinical trial (RCT) investigators to estimate the potential subject population. However, this method can lack alignment with actual conditions, potentially leading to recruitment challenges during study execution. Opening and maintaining a clinical trial site requires substantial resources, and keeping sites active with minimal or no recruitment is highly inefficient. In addition, regulatory authorities may require pharmaceutical companies to conduct RCTs within specific populations, yet recruiting an adequate sample size within a reasonable timeframe may not be feasible. It is therefore essential to present regulators with appropriate supporting evidence.

Real world data (RWD), although sometimes viewed as the opposite from RCTs, offers valuable insights that can address both of these challenges. By applying planned inclusion and exclusion criteria to RWD, it becomes possible to estimate the attainable sample size using empirical data, rather than relying solely on investigators' projections. Criteria can also be adjusted dynamically to evaluate how sample size varies, enabling identification of optimal parameters that balance feasibility with clinical relevance. Additionally, sample size estimates derived from RWD can be presented to regulatory bodies to demonstrate when conducting an RCT in a specified population is impractical. Despite these significant advantages, translating RCT eligibility criteria into operational business rules for use with RWD involves considerable complexity. In this paper, we use real-project examples (with disease and treatment details masked) to illustrate practical considerations from the perspective of a programming team.

METHODOLOGY

AGE

To begin with, age criteria seems straightforward to establish. In RCTs, age is typically determined at the time of screening or randomization. However, in real-world data, where only the year of birth may be available in the database, it is crucial to specify the exact point at which patient age should be assessed. For our project, we collaborated with the scientist who designed the criteria and determined that patient age would be defined as of the most recent diagnosis of interest.

Example:

- RCT criteria
  - Patients at least 18 years old
- RWD proxy
  - Patients at least 18 years old as of their last T2DM diagnosis on/before 31-Dec-2024

DISEASE / TREATMENT OF INTEREST

The next step involves identifying patients diagnosed with the disease of interest. In RCTs, clinical terms are frequently used, and the determination of diagnosis is typically at the physician's discretion. For our project, however, we relied on a claims database, where only diagnosis billing codes (ICD-9, ICD-10) were available. As a result, collaboration between scientists and clinical experts was necessary to translate the clinical terms from RCT criteria into corresponding ICD-9/10 codes for implementation by the programming team. To minimize the risk of administrative errors and ensure accurate identification of true cases, inclusion criteria required either at least two outpatient diagnoses separated by at least 30 days, or one inpatient diagnosis of interest during the study period.

Example

- RCT criteria
  - Patients with T2DM

- RWD proxy
    - Patients with at least two outpatient diagnoses (at least 30 days apart) or one inpatient diagnosis of T2DM between 2020-2024.
    - T2DM diagnosis codes:
        - ICD-9: 250.*
        - ICD-10: E11.*

Additionally, specific criteria must be met for patients to qualify for inclusion based on the receipt of certain treatments. In parallel, billing codes such as NDC, HCPCS, and CPT-4 are commonly utilized to accurately identify treatments.

LAB VALUES

Now we come to a tricky part: lab values. Lab values are an important aspect of RCTs because they help ensure that the patient is safe to participate in the trial. Lab values help identify underlying health issues or risks that could affect the patient's response to the treatment or their overall safety during the trial. In our project, specific laboratory criteria are used to exclude certain patients from the study cohort. Since claims data typically lacks laboratory information, it is necessary either to access a claims database linkable to lab database or to utilize an electronic health record (EHR) database, which generally includes such data. When available, LOINC codes can be leveraged to accurately identify relevant laboratory tests. On top of that, keyword searches can also be performed on test names, followed by additional data cleaning. After the pertinent records are identified, standardizing the units and converting values as required is crucial prior to applying exclusion thresholds, given the variability in reporting units. The process of handling laboratory values often entails substantial exploration and iterative communication between scientists and programming teams to optimize data quality.

Example:

- RCT criteria
    - Patients with Total IgE $\geq$ 30 KU/L *

    * The cut off is arbitrary
- RWD proxy (The steps are simplified and only for illustration purposes only)
    - Identify the IgE test
        - Use LOINC code 19113-0 or 2462-0
        - Also check using keyword search, including keywords such as: 'IMMUNOGLOBULIN E TOTAL(EXT)', 'BKR IGE TOTAL', 'TOTAL IMMUNOGLOBULIN IGE', etc
    - Remove invalid units, ex. %?%
    - Convert 'U/L' to 'KU/L' by multiplying by 1/1000
    - Remove outliers (ex. values below 0.001 or above 100,000)
    -

LIMITATION

We have discussed the wealth of information that can be obtained from RWD; however, it is important to recognize potential limitations.Certain RCT criteria cannot be directly identified within RWD. For instance, fertility data are unavailable—although age may be used as a proxy, concrete information about fertility is not captured in RWD. Similarly, severity scores such as Patient Global Activity (PGA) or the Child Health Assessment Questionnaire are seldom recorded. While they might exist in select EHRs, these datasets tend to be small and lack representativeness.

Beyond missing information in RWD, there are additional limitations to consider when using RWD to estimate potential RCT sample sizes. Although the database encompasses a large population, it does not constitute census data; the age, gender, and insured proportions may not align with those of the broader population. Therefore, comparing the database with census data is essential when projecting sample sizes to a larger population. Sensitivity analyses can also be employed to demonstrate the range of possible sample sizes, thereby enhancing the reliability of findings.

CONCLUSION

In summary, leveraging RWD to support clinical trial recruitment offers a pragmatic and data-driven approach to evaluating study feasibility, optimizing inclusion and exclusion criteria, and providing robust evidence to regulatory authorities. While RWD enables more accurate estimations of the eligible population and helps address practical challenges in site selection and recruitment, it is essential to recognize its limitations, including gaps in certain clinical details and representativeness. Collaborative efforts among scientists, clinicians, and programming teams are crucial to navigate data complexities and ensure the quality and relevance of findings. Ultimately, integrating RWD into the clinical trial planning process can enhance decision-making, reduce inefficiencies, and improve the likelihood of successful trial execution.

CONTACT INFORMATION

Yi-Chien Lee
Pfizer Inc.
Yi-Chien.Lee@pfizer.com