November 2025

Identifying rare disease patient populations via integrated claims data (US)

Background

Rare diseases are increasingly driving drug development due to scientific advancements, government incentives, and a growing understanding of genetic predisposition. According to a study published in *Nature*, an estimated 263 to 446 million people worldwide are living with a rare disease today.[1] Given the complexity of diagnosing rare disease patients, identifying real-world prevalence rates is difficult, as ranges from academic literature are often wide with varying methodologies. This can present unique challenges to biopharma companies looking to better understand the epidemiology and addressable market opportunity of a rare disease population to inform drug development and launch strategies.

Challenges with claims data

US real-world claims data can be a rich source of information for calculating prevalence rates and estimating the size of rare disease markets; such an effort, however, is not without its challenges. The following highlights three key obstacles that require careful planning and stakeholder collaboration to address:

1. ICD-10-CM diagnosis codes are often insufficient to describe rare disease patients. Rare diseases commonly have no unique diagnosis code, and the current code set lacks the specificity to describe these patients accurately.

---

[1] https://www.nature.com/articles/s41431-019-0508-0

2. The diagnostic pathway of rare disease often involves a multidisciplinary team and can include more advanced testing methods, such as genomics or other diagnostic tools. Characterizing pathways for early identification can be complex and therefore difficult to map in claims data.
3. There is a significant delay from the onset of symptoms to diagnosis—according to the National Institutes of Health, on average it takes four to five years for a patient to be diagnosed with a rare disease.[2] Because time to diagnosis can sometimes be closer to 10 years, ensuring patients are stable within a dataset for a sufficient period can be challenging.[3]

Due to the lack of coding, complexity of diagnostic pathways, and long time to diagnosis, the study design is critical when using integrated claims data in rare disease. Stitching together disparate clinical or demographic data at the patient level can be a powerful method to fill the gaps in a patient's diagnostic profile.

Data integration

Claims data, while robust and providing the most stable view of a patient over a long period, may be missing key clinical features relevant to diagnosing rare disease patients. Having a comprehensive view of a patient's diagnostic and treatment pathway can bolster the accuracy of a patient search, thus reducing false positives. This is particularly impactful with machine learning models, which benefit from having a robust set of features to train on, ultimately enhancing their differential diagnosis capabilities.

Rich datasets such as electronic medical records (EMRs), laboratory results, and genomic data can potentially fill the gaps in the patient journey and offer additional features for measurement. Rare disease patient registries are another evolving source of data that can be valuable in identifying proof-positive patients. Depending on the data available, these patients can be leveraged for look-alike modeling, which uses specific pathway and demographic characteristics to flag likely patients within a broader eligible population.

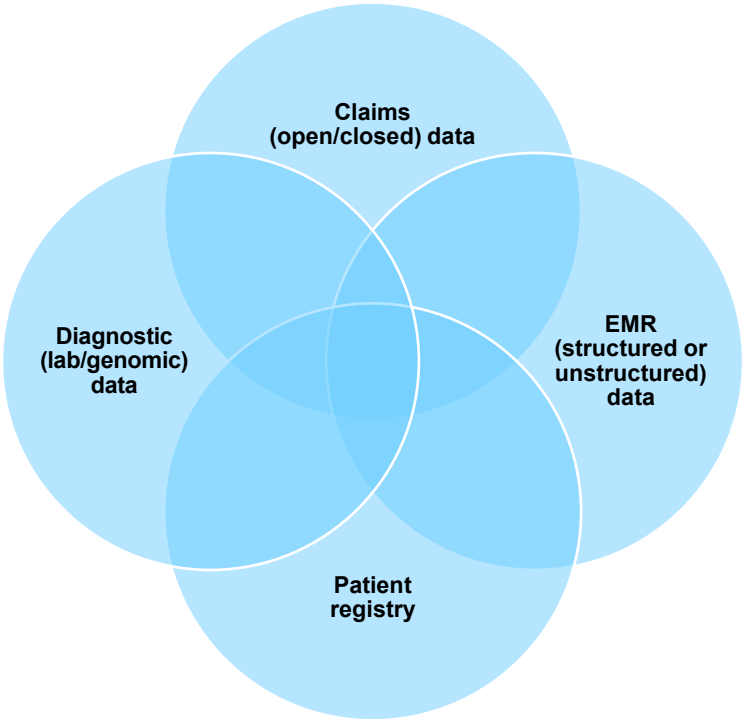
Linking claims to third-party data can allow for integrated, fit-for-purpose datasets with the relevant time, demographic, and clinical features to model a target population. The following is a general framework for integrating disparate datasets.

---

2 https://pmc.ncbi.nlm.nih.gov/articles/PMC11323401/#CR4

3 https://rarediseases.org/center-of-excellence/

Figure 1: Data overlap framework



While there is a strong value proposition for using integrated data to identify rare disease patients, execution can be challenging without alignment on key endpoints and the right methodology to match. Common shortfalls include:

- Overlapping cohorts too small to be studied;
- Inability to meet patient stability or continuous enrollment criteria; and
- Fill rates of relevant features not robust enough to measure.

In addition to data quality, navigating the complex contracting environment across different data providers can create obstacles at the onset of a project. Privacy-Preserving Record Linkage (PPRL), also known as “tokenization,” is a method of linking patient records across disparate datasets. Companies that have deployed this method commercially across the healthcare data landscape (e.g., Datavant) can potentially be leveraged to circumvent challenges related to linking disparate datasets at the patient level.

There is no one-size-fits-all approach for using integrated data to identify rare disease patient populations. Every project requires a clear articulation of project outcomes, careful planning across internal stakeholders, and partnership/collaboration with third-party organizations.

Framework for success

To address the challenges associated with integrating multiple datasets, we suggest a multiphase approach to assist biopharma organizations in measuring the real-world prevalence of rare disease populations:

Phase 1 Perform data feasibility and overlap analysis

Keys to success include:

- Multi-stakeholder involvement early on
- Alignment on short-term/long-term objectives
- Phase informed by applicable market research

Phase 2 Develop analysis plan

Keys to success include:

- Well-defined primary/secondary/exploratory endpoints
- Patient qualification rules aligned with qualitative research
- Methodology to control confounders customized based on results from Phase 1

Phase 3 Execute analysis plan

Keys to success include:

- Interim readouts at critical milestones
- Revised methodology as needed to optimize results
- Inclusion of internal stakeholders to ensure data outputs align with messaging

Starting with this general framework for sizing difficult-to-identify markets and then customizing it to meet specific objectives can facilitate the use of integrated data in rare disease patient populations to inform demand forecasts, clinical trial design, and commercial launch planning.

If you need assistance with and/or would like to discuss how your organization is thinking about leveraging data to identify rare disease patients, please contact Cliff Li, principal in CRA's Life Sciences Practice.

About CRA's Life Sciences Practice

The CRA Life Sciences Practice works with leading biotech, medical device, and pharmaceutical companies; law firms; regulatory agencies; and national and international industry associations. We provide the analytical expertise and industry experience needed to address our clients' toughest issues. We have a reputation for rigorous and innovative analysis, careful attention to detail, and the ability to work effectively as part of a wider team of advisers. To learn more visit **www.crai.com/lifesciences**.

Contact

Cliff Li
Principal
Chicago
+1-847-209-1755 (mobile)
cli@crai.com