The Importance of Integrated Registries in Rare Disease Research

There is a clear need to integrate and harmonise data from existing registries to better understand disease natural history and to enhance the success in developing new treatments to counter the impact of rare diseases. How can registries be more effectively integrated?

Dr Annette Stemhagen, Irene Cosmatos, and Aaron Berger at UBC

Registries have long played a vital role in biopharmaceutical research and development by providing insight into disease natural history, real world patterns of care, and safety and effectiveness of treatments. In 2021, both EMA and FDA released guidance outlining a regulatory framework for utilising data from registries in regulatory decision-making and for the implementation of post-marketing commitments (1).

Registries are of paramount importance for generating insights that can lead to improved patient care in rare diseases, often through a better understanding of the underlying biological mechanisms and genotype-phenotype associations. The small number of cases for each disease, however, prevents the creation of large cohorts necessary to derive meaningful analysis. Consider that rare disease in the US is defined as a cumulative number (impacting 200,000 or fewer people; the US population is approximately 330 million people), while the EU defines a rare disease as one that affects less than five per 10,000 people.

Often in rare disease there exists a fragmented global landscape of registries and 'siloed' data collection initiatives. These initiatives are generally developed as independent patient registries or even investigator-initiated cohorts from a single practice, and are often restricted to patients from one particular geographic region or country. These initiatives often lack a uniform data collection format that would easily enable data sharing. Furthermore, launching a new registry is a costly endeavour, and may introduce a duplicate data entry burden on healthcare providers if a registry already exists in a given disease.

Many rare diseases with underlying genetic etiology have distinct clinical subtypes (e.g., juvenile vs infantile onset) that differ in phenotypes as well as effective treatments. Integrated data strategies are important in rare disease, especially to enable subgroup-focused analyses. By bringing together data from multiple registries, a larger database is available from which subgroups of interest can be studied. This approach is further enhanced by enriching combined datasets through linkage with other, more specialised secondary data (e.g., mortality data) to increase the value of the data for disease natural history understanding.

These dynamics highlight the need for integrating and harmonising data from existing registries to amplify the generation of meaningful data, which are critical in understanding the disease natural history and to contextualise the safety and effectiveness of new treatments for patients suffering from rare diseases.

Building an Integrated Registry

Gap analysis and a common data model

One of the most significant challenges to integrating diverse registries is to develop the strategy for standardising data elements into a common data structure. This common data model represents a structure for consolidating data elements across registries that are using different data collection tools, and establishes a common vocabulary to capture important data elements such as diagnoses, medications, procedures, and genetic and laboratory results. A challenge for data integration is that not all registries use the same data collection tools or collect the same information. Prior to creating the common data model, a robust gap analysis is required to organise the data variables according to categories that include: identical, comparable, and missing data elements.

Patient Identity Management

A robust patient identity resolution approach is a lynchpin for any successful data integration project. This starts

with assigning a persistent and unique 'identifier' or 'token' to each patient using a sophisticated matching algorithm. This technology enables downstream functions in the registry integration project – such as deduplication and linkage of data across different datasets in a privacy-compliant manner – without the need to share Personal Identifiable Information between parties.

Patient privacy is a primary concern. Integrating patient information presents challenges of de-identification. A unique patient identifier can be assigned to each patient, which enables matching across the 'siloed' independent disease registries – as well as with various secondary data sources. This approach maintains compliance with the Health Insurance Portability and Accountability Act, General Data Protection Regulation, and the California Consumer Protection Act requirements. Furthermore, a careful review of the informed consent forms being used by each existing registry is important for gaining an understanding of the permitted uses of patient-level data.

Deduplication of patients is a critical capability in creating a research-quality integrated registry. When the data pool is small, as in rare diseases, there is opportunity for patients to be enrolled in more than one registry, especially when the registries have different research agendas (e.g., genotype-phenotype vs. treatment effectiveness). Additionally, patients may participate in registries through multiple doctors they have seen over time or through different foundations. Every attempt must be made to identify patient data that appear in more than one registry, otherwise the results may be skewed and conclusions erroneous.

Data Enrichment

In some situations, linkage to secondary data (e.g., pharmacy claims, electronic medical records, device data, lab data, mortality data, consumer data) may be a desirable approach to fill in gaps in evidence, or to create a richer view of the patient journey. These data, in many cases, can be combined with existing registry data to fill in missing information (i.e., data not collected uniformly across the registries), as well as help provide a more complete picture of patient symptoms – both prior to and post-diagnosis.

Linking secondary data can also help create a more longitudinal view of the journey of patients with rare diseases. Incomplete follow-up, patient mobility, attrition, and other factors can damage the completeness of a data set. By integrating secondary data, as well as lengthening the registries with prospective data, researchers can gain far more insight into small patient populations.

Stakeholder Collaboration

What is the value proposition for registries and consortia to participate in data integration initiatives? No one dataset by itself is sufficient. However, by expanding the research to include academic and community practices – urban and rural settings on a global basis – researchers are able to achieve a multi-dimensional view of the patient journey.

Physicians treating patients with rare diseases are altruistically motivated to learn more about potentially beneficial treatment approaches. It is important to plan for sharing the results of the data integration initiative with the physicians and patient and caregiver community.

Summary

Integrated registries play a vitally important role in developing an understanding of disease natural history, real-world patterns of care and safety, and effective outcomes of treatments. Registries become much more powerful when they are combined, particularly in rare diseases where patient populations are small. Effectively integrating registries requires a combination of data science and technical expertise to support this kind of RWE generation.

Reference

1. *FDA: Assessing Registries to Support Regulatory Decision Making for Drug and Biologic Products, Nov 2021; EMA: Guideline on Registry Based Studies (EMA426390/2021: 22 Oct 2021*

Dr Annette Stemhagen, Senior Vice President and Chief Scientific Officer at **UBC**, has more than 30 years of public health research experience, and provides strategic consultative services to pharmaceutical, biotech, and medical device clients. Dr Stemhagen is active in the International Society for Pharmacoepidemiology and the Drug Information Association. In 2004, Dr Stemhagen was appointed as the first Industry Representative to the FDA Drug Safety and Risk Management Advisory Committee.

Irene Cosmatos, MSc, Senior Research Specialist, RWD Epidemiologist at **UBC**, is an experienced epidemiology database analyst who supports product training and development within UBC's Database Analytics Automation practice area. She leads UBC's retrospective database analytic service activites, which include conducting global database feasability evaluations, developing appropriate database study designs, and overseeing the execution of database studies.

Aaron Berger, PMP, Executive Director US Late Stage Operations and Real World Evidence at **UBC,** has served a diverse range of roles leveraging global capabilities in the execution of pre- and post-marketing product development programs and real-world evidence generation. He has supported the develeopment of RWD/RWE architectures and solutions to drive greater insight into the safety profile and value proposition of medicines.