What Can Disease-Specific Data Sources Offer Us?

Marta Pereira, PhD, Research Associate, Real-World Evidence, Evidera
Dimitra Lambrelli, PhD, Research Scientist, Real-World Evidence, Evidera
Sreeram Ramagopalan, PhD, Senior Research Associate, Real-World Evidence, Evidera

Introduction

The importance of real-world data both in drug development and post product launch is well known. However, what needs increased recognition is that the quality and ultimate utility of a study using real-world data is highly dependent on the information available in the data source. Commonly used, real-world data sources include electronic medical records and administrative health insurance claims data, but in a world where there are often many therapeutic options for a disorder, there is a pressing need to obtain detailed data on a disease in order to provide insights into unmet need and treatment effectiveness. As such, there is a shift occurring from using routinely collected data to disease-specific data sources.

What are disease-specific data sources?

Disease-specific data sources are databases, registries, or studies using observational methods to evaluate a specific population of people with a disease. Standardized information is collected about the patients, and it may be cross-sectional or longitudinal in nature. Disease-specific data sources are generated for the purpose of observational data collection that can be used for a specific research agenda, including the monitoring of disease natural and treated history, patient outcomes, and the study of best practices in care or treatment. They may pursue a specific, focused research agenda, collecting data for a limited time to answer a specific research question, or may collect data on an indefinite basis to answer a variety of existing and emerging research questions. The sources may be organized and operated in a variety of forms and formats.

"...there is a shift occurring from using routinely collected data to disease-specific data sources."

In this article, we will explore the advantages and disadvantages of using disease-specific data sources in real-word studies, using two case studies; multiple sclerosis, a specific neurological disease with a recent sea change occurring in disease treatment; and cancer, a highly prevalent disease appearing in many different forms and a corresponding vast therapeutic armamentarium.

Case Study - Multiple Sclerosis

Multiple sclerosis (MS) is the most common neurological disorder affecting young adults in North America and Europe. About 85% of patients present initially with relapsing-remitting MS (RRMS), characterized by recurrent episodes of neurological dysfunction interspersed with periods of lack of apparent disease activity.[1] At present, there are 13 disease-modifying therapies (DMTs) approved by the U.S. Food and Drug Administration (FDA)[2] and 11 DMTs approved by the European Medicines Agency (EMA)[3] for the treatment of RRMS, with new treatment options emerging each year.

The clinical trials that led to the approval of these treatments for RRMS are recognized for their limitations in terms of providing data on efficacy rather than effectiveness. They further lack the ability to provide more general epidemiologic data about MS, for example disease incidence; most frequent reasons for hospitalization in patients; and major drivers of cost of care.

Routinely collected data, that is administrative data collected by insurance companies (e.g., sickness fund data from Germany) or electronic medical records (e.g., the Clinical Practice Research Datalink [CPRD] in the United Kingdom), provide the mainstay for real-world

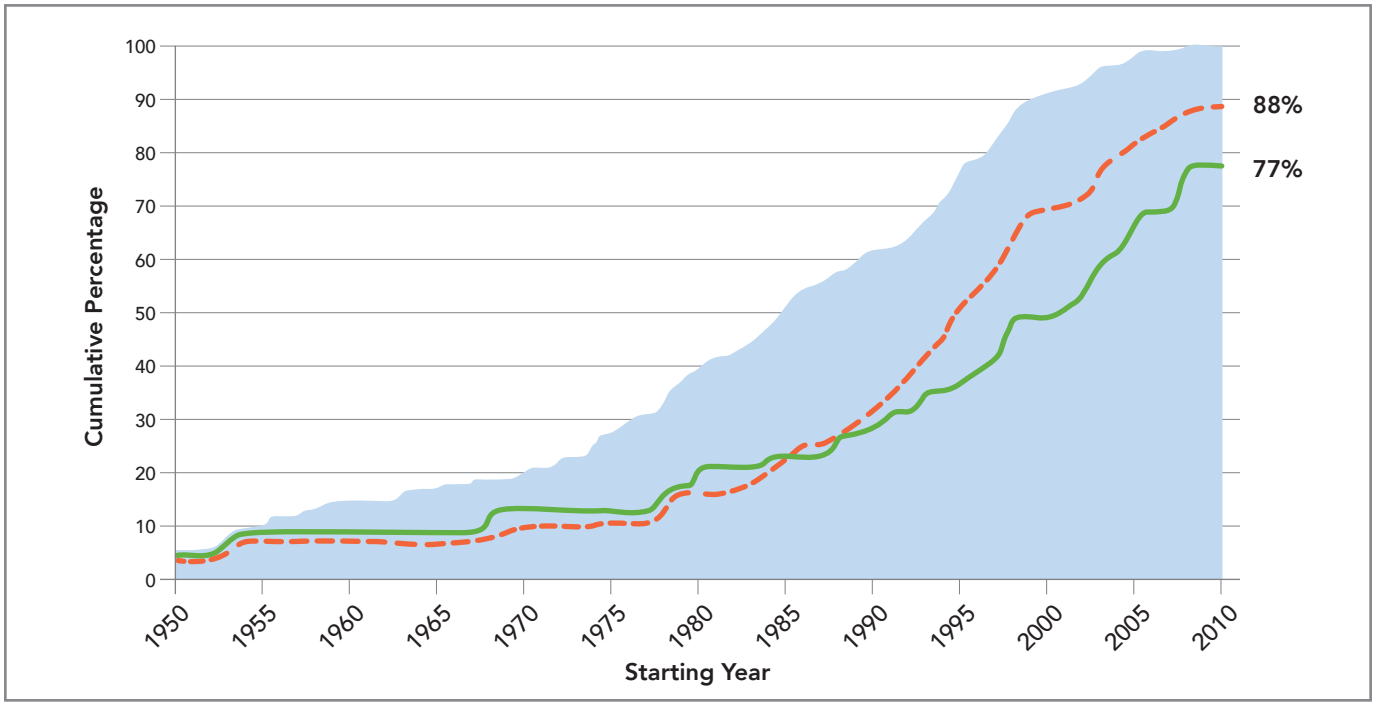
data analyses. There are a number of benefits to using such data - namely they are readily available and relatively inexpensive. However, there are a number of limitations with routinely collected data. In Europe there is a lack of good quality and sufficiently representative data in many countries. For data sources that do exist, their biggest deficiency is the incompleteness of data. Data sources generally do not include different types of care - they may be focused on primary care or the hospital sector, but rarely cover all the different settings that play a role in medical treatment. This is becoming more important for MS as the availability of newer monoclonal antibody therapies increases the number of treatments given in secondary care, which is not captured in primary care medical record data such as CPRD. Routinely collected data also lack clinical detail, for example information on disease severity measures such as the widely used Expanded Disability Status Scale (EDSS) or magnetic resonance imaging (MRI) results. The coding for identifying patients generally will not allow patients with different forms of MS to be distinguished. Clinical data are critical for analyses of patient outcomes and are a key determinant of prescribed treatment, so missing this information restricts the analyses that can be done.

One way of obtaining more in-depth data on patients with MS is to use a disease-specific data source. Examples of disease-specific data sources in MS include the European Database for Multiple Sclerosis (EDMUS), the Swedish National MS Registry, the Danish National MS Registry, and the global MS Registry (MSBase).[4]

"Data sources generally do not include different types of care - they may be focused on primary care or the hospital sector, but rarely cover all the different settings that play a role in medical treatment."

Looking at one of these data sources provides an illustration of the data available in comparison to general population data sources. MSBase is a longitudinal online registry and is open to any neurologist worldwide to collect data on MS patients. It is registered as a not-for-profit organization in Australia. To initially register a patient on the database, a minimum set of data are needed that encompasses MS course, diagnosis date, EDSS score, paraclinical tests (e.g., MRI), relapse dates, and treatment dates. Further, data entry for at least an annual follow-up visit is required for each patient.[5] Clearly, therefore, this data source fills some, but not all, of the clinical and treatment data gaps present in routinely collected data.

Figure 1: All General Cancer Registries Combined (n=99)



Starting year of the CR and cumulative rate of recording of stage (dotted red line) and treatment data (solid/green line) in population-based cancer registries in Europe ordered geographically.

Figure 2: Uses for Cancer Registry Data

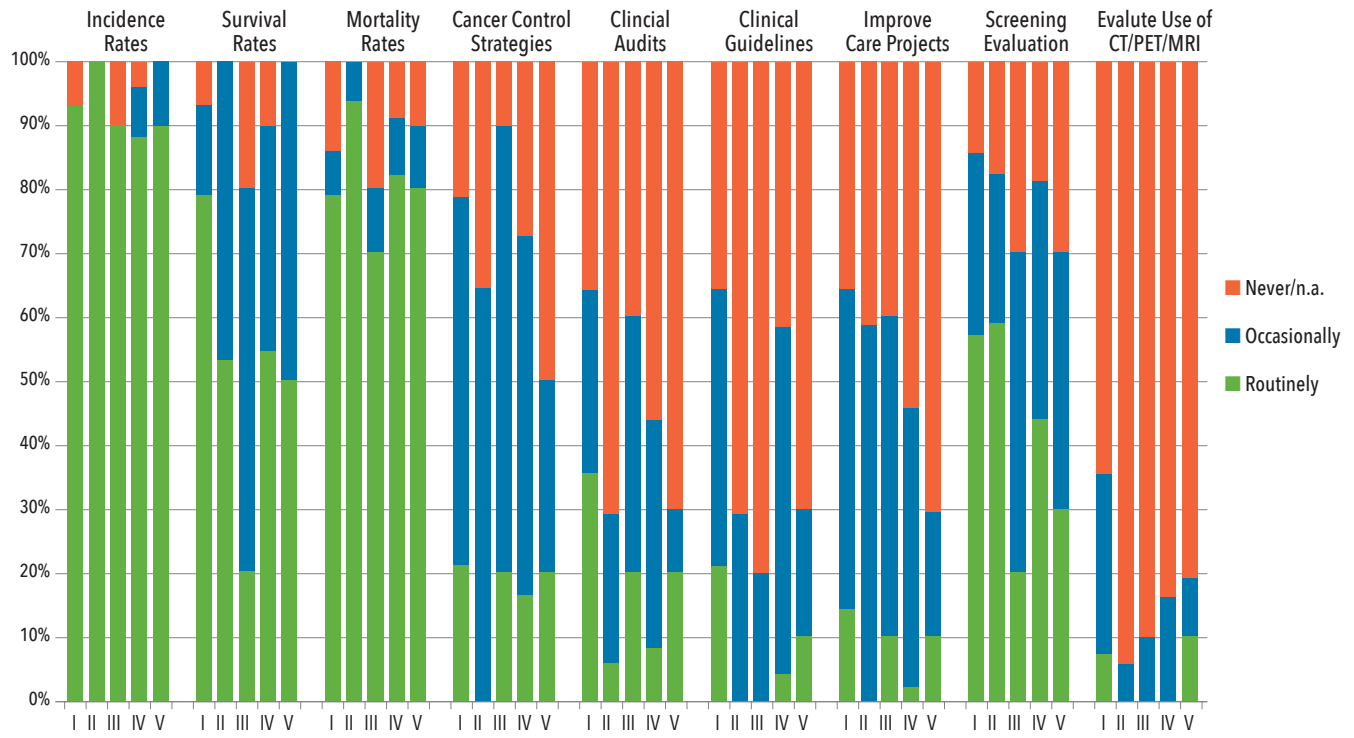


Fig. 2 Use of PBCR data (*n*=99) for estimations of incidence and mortality for cancer control strategies, clinical audits, clinical guidelines, care improvement care projects, mass screenings evaluation and utilisation of computerised tomography/positron emission tomography/magnetic resonance imaging (CT/PET/MRI) according to European sub-region.

Case Study – Cancer

Cancer is a major public health problem in the United States (U.S.), in Europe and in many other parts of the world. It is currently the second leading cause of death in the U.S.[6] and in Europe[7] after cardiovascular disease.

In the area of oncology, routinely collected data by insurance companies or electronic medical records similar to CPRD, despite some advantages, have certain limitations, which as for MS are mainly related to the absence of certain data. The most important data gaps are:

- **clinical indicators:** e.g., stage, ECOG (electrocorticography), histology, cytology, morphology
- **medical treatment:** as oncology treatment is mostly hospital-administered, data on treatments are absent from the majority of general datasets. The reasons for this is either due to the fragmented nature of the datasets (e.g., covering only primary care) or the application of DRG systems that would not allow the identification of individual drugs given within the hospital setting. Subsequently any information about the duration of treatment, treatment cycle, reasons for treatment discontinuation, and response to treatment is also absent.
- **adverse events:** due to the coding system used for diagnosis of certain conditions (i.e., usually International Classification of Diseases - 9 or 10), specific adverse events are not appropriately recorded (e.g., nausea, vomiting, or other probable adverse events without a respective ICD diagnosis code).

The limitations that stem from the non-availability of certain variables affect all types of studies in the area of oncology: treatment pattern, resource utilization, and burden of illness studies. Epidemiological studies are also affected. Over recent years, population-based data are increasingly used to estimate survival in different cancer populations. However, survival reflects not only treatment but also prognostic factors, such as stage at diagnosis, histological type, and other characteristics of the disease. In the absence of these factors, the reasons for any variations in survival observed cannot be properly identified. Moreover, in the area of oncology the value equation for an oncology product may also be enhanced by demonstrating the impact of therapy in specific patient subgroups, for example non-responding patients. Looking at the list of recently approved haematology oncology drugs by the FDA, it becomes evident that such evidence is convincing to regulators and payers who seem willing to offer therapies to the patient

subpopulations most likely to benefit from therapy. Therefore, the availability of data on medical treatments and key clinical oncology indicators is paramount in the analysis of patient outcomes.

Cancer Registries

Cancer registries have a long history with the first attempts made in the early 1900s in different countries to estimate the number of new and existing cancer cases in given populations. Cancer registries can be grouped into three types: 1) facility-specific registries that collect information about patients diagnosed and treated at a specific facility; 2) specialty registries that only collect information on specific types of cancer (e.g., paediatric cancers); and 3) central cancer registries that collect information about cancer patients in a specific geographic area (country, region, etc.).[8] The main purpose of existing registries is to develop intelligence to monitor and drive improvement in prevention, standards of cancer care, and clinical outcomes of cancer patients. However, the intended purpose of each registry might differ, and this is what defines the necessary properties of the data to be collected. The Cancer Outcomes and Services Dataset (COSD) in the United Kingdom is the national dataset for reporting on cancer in the National Health Service in England. The dataset includes comprehensive data on patient demographic characteristics but also cancer specific data, i.e., morphology, cytology, white blood cell count, platelet count, performance status, whether a patient participates in a clinical trial, tumour-node-metastasis staging classification procedures performed, and patient death details. In addition to core data collected, the dataset also includes cancer-type-specific information.

"Real-world data in its various forms – … have an important role to play in the evaluation of epidemiology and burden of disease, treatment patterns, compliance, persistence, and health outcomes of different treatments."

For example for breast cancer, additional information is collected on clinical assessment results, mammogram results, Nottingham Prognostic Index (NPI) Score, ASA score, invasive grade, tumour size, Human Epidermal Growth Factor Receptor 2 status, cytology, and biopsy results. Although data on specific treatments received are incomplete, these data provide a wealth of cancer-specific information that is lacking from administrative datasets. However, the nature of each cancer database may vary significantly. The Swedish cancer registry for example, provides detailed information on cancer incidence, mortality and prevalence. The site of the tumour, histological type, basis and date of diagnosis, and stage are being collected at an individual level along with information on patient's death. However, further clinical and treatment data are not available. The amount of data registries collect is increasing with time (see Figure 1) due to the increased awareness of the importance of such data (see Figure 2).

Table 1: Advantages and Disadvantages of Disease-Specific Data Sources

Advantages	Disadvantages
Focused study population	Limited study population – generalizability only for that specific group of patients
Specific and detailed clinical information: • disease severity measures • disease specific treatment	No comparator group (e.g., individuals without disease or with other disease[s])
Study design selected according to the natural history of the disease (e.g., time between follow-up evaluations)	Differences between different datasets. Feasibility assessment required for data content and quality
Patient subgroups analysis based on various disease-specific indicators	Cost data is not available
Only feasible method to study patients with rare diseases	Certain registries do not collect information on treatment pathways that are not disease- specific
Follow-up of patients during their entire treatment pathway	

Conclusion

In summary, disease-specific data sources are available for certain diseases and provide solutions to data gaps in more administrative type datasets. Nevertheless, disease-specific data sources do have their own drawbacks (which vary depending on the source and the methodology used to collect data) and these must be borne in mind. Table 1 provides a summary of key advantages and disadvantages. These shortcomings include the fact that for some sources, the data represent a restricted study population – generalizable only for that specific group of patients. Further, information available is limited to what was collected, so data may not serve a wider range of research purposes. Finally there may not be an appropriate comparator group within the data (e.g., individuals without disease or with another disease).

Real-world data in its various forms – routinely collected or disease specific, longitudinal or cross-sectional, retrospective or prospective - have an important role to play in the evaluation of epidemiology and burden of disease, treatment patterns, compliance, persistence, and health outcomes of different treatments. Many study designs are possible but the limitations of each require careful consideration. A critical assessment of the available data sources to identify those that yield the best information for the study needs is essential. An informed decision must take into account several characteristics of the data source, including data content and data source accessibility.

For more information, please contact Marta.Pereira@evidera.com, Dimitra.Lambrelli@evidera.com, or Sreeram.Ramagopalan@evidera.com.

REFERENCES

[1] Ramagopalan S, Dobson R, Meier UC, Giovannoni G. Multiple Sclerosis: Risk Factors, Prodromes, and Potential Causal Pathways. The Lancet. 2010; 9(7):727-739.

[2] National Multiple Sclerosis Society – Medications. Available at: http://www.nationalmssociety.org/Treating-MS/Medications. Accessed September 18, 2015.

[3] European Medicines Agency. Available at: http://www.ema.europa.eu/.

[4] Eraksoy M, Butzkueven, Ziemssen T, Zivadinov. Time for Change – Evolution of Real-world Evidence Outcome Measures in Multiple Sclerosis Exemplified by Fingolimod. *Eur Neurol Rev.* 2014; 9(2):136-142.

[5] Butzkueven H, Chapman J, Cristiano E, Grand'Maison F, Hoffmann M, Izquierdo G, Jolley D, Kappos L, Leist T, Pöhlau D, Rivera V, Trojano M, Verheul F, Malkowski JP. MSBase: An International, Online Registry and Platform for Collaborative Outcomes Research in Multiple Sclerosis. *Mult Scler.* 2006 Dec; 12(6):769-774.

[6] Centers for Disease Control and Prevention. Leading Causes of Death. Available at: http://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm. Accessed September 21, 2015.

[7] Eurostat. Causes of Death Statistics. Available at: http://ec.europa.eu/eurostat/statistics-explained/index.php/Causes_of_death_statistics. Accessed September 21, 2015.

[8] Zachary I, Boren SA, Simoes E, Jackson-Thompson J, Davis W, Hicks L. Information Management in Cancer Registries: Evaluating the Needs for Cancer Data Collection and Cancer Research. *Online J Public Health Inform.* 2015; 7(2):213.