Epiphany

Enhancing Cancer Epidemiology: Why Multiple Data Sources Matter

Epiphany Partners Inc., June 2025



Introduction

Cancer and rare disease research are evolving fast as pharmaceutical companies race to bring more precise, personalized treatments to market — but one critical piece often lags behind: epidemiology data. Or more specifically, the way we estimate and utilize the patient numbers.

Given the lack of highly rigorous cancer epidemiology across the globe, it is often tempting to rely on a single data source, especially if it's been used historically. It's easy, it's simple, it's already in the forecast systems, but relying on one epidemiological dataset is like trying to guess the puzzle from only one piece. You miss nuance; you miss context; you might even miss the market.

For pharma leaders across R&D, clinical development, and commercial strategy, it's time to rethink the epidemiology playbook. In this article, Epiphany Partners unpacks why multiple data sources are essential for robust cancer and rare disease assessments — and how to make them work for you.

The Problem with One-Stop Data Shops

Let's get right to the point: **single-source epidemiology data often falls short**. Depending on the data source, each data set has limitations in terms of scope, methodology, data availability, and often lacks details relevant to the complex landscape of modern oncology and rare diseases.

Here's why:

1. Incomplete Population Coverage

- Many sources are limited in geographic coverage
- Many underrepresent rural, minority, or uninsured populations
- Detailed data by subtype, histology, biomarker, and other demographic issues is often limited across multiple sources
- Rare diseases usually have little to no available data, with most rare diseases not included in population-based registries

2. Lack of Transparency

- Methodologies are often proprietary and are presented as "black box" functions without any details or assumptions to allow for comprehensive assessments
- Input data sources may not be disclosed or may include additional assumptions that are not implicitly stated
- You might not know how incidence, prevalence, or treated patient estimates were derived, including all inputs and assumptions

3. Rigor and Depth of Data

- Limited available details on subtypes/histologies, stages, and molecular markers
- Inability to assess the rigorousness of the patient flow process
- Little to no data on comorbidities or treatment histories, given the need for additional data inputs/sources

Bottom line: If you're basing key strategic or investment decisions on a single epidemiology source, you could be building the commercial foundation on sand.

The Types of Available Epidemiology Data

There are several different types of available epidemiology data, each with their own strengths and weaknesses when taken as a single source.

1. Internal Company Epidemiology

- **Strength:** Internal datasets are usually focused on existing brand indications
- **Limitations:** Often lack depth and breadth of coverage outside of existing brand indications, and there may be uncertainties about data updates and ability to expand coverage by cancer, subtype, biomarker, and/or countries

2. *EpicOncology*™ Cancer Epidemiology Dataset

- **Strength:** Provides broad coverage for both diseases (42 cancers) and countries/regions (42 countries), focus on country-specific datasets for key inputs (incidence, mortality, survival, staging), robust patient flow analyses and modeling, transparency in terms of sources, assumptions, and methodology, consistency of data across all countries for each study canceer
- **Limitations:** Treatment data is based on physician surveys (U.S., Europe, and Japan) and may not be specific for each country

3. Other Commercial, Epidemiology Data Sources (Clarivate, Norstella, Oracle)

- **Strength:** Usually provide broad coverage for both diseases and geographic regions/countries (with a focus on the G8 countries: U.S., Japan, China, and EU5 (France, Germany, Italy, Spain, United Kingdom)
- **Limitations:** May lack necessary treatment detail for forecasting and modeling, have transparency issues for inputs and assumptions, may have limited methodology detail, and lack consistency of coverage across all cancers and/or countries

4. Public Epidemiology: Global and Country-Specific

- **Strength:** Usually broad general coverage by countries across major cancer types
- **Limitations:** Often limited by smaller disease states and subtypes, including biomarkers, most often lack staging data (necessary for forecasting), usually lack robust forecast estimates, and usually provide only incidence and/or prevalence estimates

5. Peer-Reviewed Epidemiology Literature

- **Strength:** Most recent datasets using accepted methodologies and clear assumptions with robust reviews prior to publication
- **Limitations:** Usually highly limited in scope and coverage (usually one cancer or subtype) with limited ability to forecast estimates or apply to broader populations

Understanding the Complexities of Multiple Sources

Epiphany Partners, as one of the leaders in cancer epidemiology consulting, is often asked about epidemiology sources, limitations of available data, and comparisons between different data sets. Brand and development teams are trying to better understand their target markets but are often limited by the epidemiology data.

Unfortunately, the ability to address these types of questions is complicated by the fact that most data sets use various source data, different methodologies, and often unclear forecast processes. Without transparency across sources, the ability to explain differences and compare strengths and weaknesses across and between data sets.

Even publicly available data sets, like GLOBOCAN and American Cancer Society's Cancer Facts and Figures, provide general equations and methodology descriptions, but not actual model data to provide clear insights into the final data.

Overall, the following are key features that you should look for in epidemiology data sources:

1. Detailed Data Sources and Assumptions

Availability of the epidemiology data sources, assumptions, and rationale for usage provides you with a baseline level of confidence in the strength of the epidemiology data.

2. Access to the Epidemiology Model

Being able to clearly understand the model organization and mathematical structures is important to provide trust and accountability in the data source. Data sets that lack direct access to the model make it harder for you to verify and validate the approaches and inputs to the model.

3. Clarity in Patient Flow Process

Accurately capturing how patients move through the treatment process from diagnosis to treatment to follow-up is critical to rigorous epidemiology data. Making sure the epidemiology data source provides an appropriate estimate of the patient journey results in more actionable basis for future shifts in the market

Having these features available allow you to confidently reference and utilize the resulting data. The lack of any of these features makes the epidemiology modeling more of a "black box" where you may be limited in understanding the prediction process and may hinder your interpretation of data trends and forecasts.

The Case for Using Multiple Epidemiology Sources

Integrating data from multiple sources might take more effort, but it pays off in the long run. It allows you to leverage strengths while minimizing limitations or weaknesses of different epidemiology sources:

1. Better Representation

Combining national cancer registries, hospital-based studies, patterns-of-use studies, and claims/real-world evidence/treatment survey data creates a fuller, more accurate picture of the disease landscape.

Example: Integrating SEER (U.S.), CONCORD (global), and hospital-based data from Asia, the collective findings from multiple sources revealed significantly higher liver cancer rates in underrepresented subpopulations — helping identify critical gaps in clinical trial coverage

2. Greater Methodological Rigor

Multiple datasets mean you can compare results and try to identify reasons or issues for differences from various sources. This ability to cross-verify is important to better understanding the marketplace. That means fewer assumptions, fewer blind spots, and more confidence in your numbers.

Think of it like this:

DATA SOURCE	STRENGTH	LIMITATION
Commercial Data	Detailed regional coverage, robust forecasted populations	Population-based, limited clinical data
GLOBOCAN	Global coverage, multiple methodologies, limited projections	Model-based, not patient-level
Real-world data	Rich treatment/pathway insights	Not population-representative
Claims databases	Broad payer insights	Limited clinical detail

No single source is perfect — but together, they can provide powerful insights and robust foundations for population forecasting.

3. Richer Data for Deeper Insights

In cancer especially, subtypes matter. The difference between HER2+ and triple-negative breast cancer isn't just academic — it drives everything from protocol design to commercial strategy. You won't get that from a high-level registry summary.

To gain insights into smaller subtypes, more rigorous datasets are required as inputs, along with clear methodological approaches, and robust forecast processes. Using multiple datasets may be necessary to generate the level of detail required for any assessment or forecast.

Methodologies Matter

Even when using the exact same datasets, different sources will generate different patient estimates, based on how the raw data is used and which forecast methodologies are selected.

The following shows the impact of three different methodologies using the same SEER/NPCR dataset for the U.S. for the estimate of "All Cancer Sites" in 2021

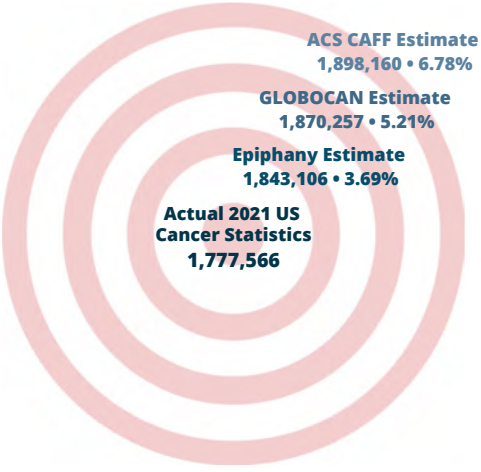


Figure 1: Comparison of Different Epidemiology Sources for 2021 U.S. All Sites Incidence Estimates, versus reported 2021 US Cancer Statistics Actual Value

1. Epiphany's *EpicOncology*™ Database

- **Methodology:** Epiphany uses a joinpoint linear regression model by aggregated age-groups and gender
- **Output:** 2021 All Sites Estimate of 1,843,106 patients, 3.69% difference over the reported 2021 U.S. Cancer Statistics Actual Incidence of 1,777,586

2. GLOBOCAN 2021 Estimate

- **Methodology:** GLOBOCAN uses a short-term prediction model applying a Poisson distribution for the U.S. (other countries may use different methodologies)
- **Output:** 2021 All Sites Estimate of 1,870,257 patients, 5.21% difference over the reported 2021 U.S. Cancer Statistics Actual Incidence of 1,777,586

3. American Cancer Society's Cancer Facts and Figures 2021

- **Methodology:** ACS uses a spatiotemporal small area model to forecast the historical SEER/NPCR to 2021
- **Output:** 2021 All Sites Estimate of 1,898,160 patients, 6.78% difference over the reported 2021 U.S. Cancer Statistics Actual Incidence of 1,777,586

The different approaches used by three different sources resulted in variations of the estimated incidence. The assumption often made is that if the original datasets are the same, the outcomes should also be similar, but clearly methodology matters when forecasting cancer and rare disease epidemiology.

Triangulation: Epidemiology's Secret Weapon

In epidemiology, triangulation refers to the use of multiple data sources, methods, or investigators to study a single phenomenon. It's a classic technique in causal inference, and in cancer epidemiology, it's invaluable.

Why Triangulation Works:

- It helps **validate findings** across different settings
- It reduces the risk of systematic error or bias
- It supports **robust patient estimates**— even when individual sources are incomplete or lacking details

As noted in the published study (Lawlor 2017), triangulation can improve the internal validity of research findings, particularly when data from randomized trials are unavailable (which is most often the case in cancer and rare disease epidemiology).

In Practice: Future Lung Cancer Estimates

The following are actual commercial epidemiology source data for Stage III/IV non-small cell lung cancer (NSCLC) in the year 2030 for the United States. Given the future focus, there is no practical way to validate the numbers, but a comparison of the data can illustrate issues between the sources and the need to consider multiple data sets and triangulation.

Statistic	Subtype	EpicOncology™	Other Source
Incidence & Newly Metastatic (Epic) vs. Newly Recurrent (Other)	Squamous Cell	42,961	65,040
	Non-Squamous Cell	122,485	152,434
First-Line Drug Treated Patients	Squamous Cell	34,702	33,689
	Non-Squamous Cell	99,846	89,570

A quick review of the data shows that the estimates for the drug treated are similar across both sources, but the incidence & metastatic/recurrent patient numbers are significantly different (the Other Source is 20 to 35% higher than the EpicOncology™ data). Given that the two sources are reporting different statistics, it is hard to directly compare the data (newly metastatic only includes patients who progress to metastatic disease, while newly recurrent could include all recurrent patients, including those who become metastatic and those who are still non-metastatic).

Source	Analysis	Stage III	Stage IV
EpicOncology™	Newly Metastatic	—	All non-metastatic patients who progress to the metastatic setting in 2030
Other Source	Newly Recurrent	All recurrent Stage III patients	All recurrent Stage IV patients (including newly metastatic patients)

Given that the first-line drug treated patients are the same, the approaches to estimating treated patients may be similar. In this context, using triangulation for the first-line drug treated could provide a fine-tuned patient estimate.

Statistic	Subtype	Triangulated Estimates	EpicOncology™	Other Source
First-Line Drug Treated Patients	Squamous Cell	34,196	34,702	33,689
	Non-Squamous Cell	94,708	99,846	89,570

This example shows that triangulation, when applied appropriately across data that are similar, can potentially enhance patient estimates. However, triangulation may be inappropriate in settings where the data are too dissimilar or where it is hard to determine if the data are representing the same parameters.

In most cancers or rare diseases, no single data source can be considered "comprehensive" for meaningful estimates since each source has made assumptions, especially future forecasts. But appropriately triangulating patient estimates across sources can provide credible, actionable patient estimate ranges.

Why This Matters for Pharma Leaders

Still wondering what all this means for your work? Here's why you should be considering using **multiple sophisticated epidemiology sources:**

If you're in Brand, Commercial, or Access Strategy:

- **Forecast markets more accurately** using triangulated incidence and treated patient numbers
- **Develop detailed subtype and target populations** using multiple sources to account for missing or incomplete patient data
- **Tailor pricing and access models** to reflect care gaps by region, subtype, and/or biomarker
- **Position your asset** as a differentiated solution to an unmet, under-quantified market, instead of just part of a large target pool

If you're in Outcomes Research, Real-World Evidence, Health Economics, or Epidemiology:

- **Triangulation to improve validity** and increase confidence in the data, whether the data converges or diverges
- **Addressing methodological issues** by referencing multiple sources to identify and quantify any biases
- **Capture the reality of patient experience** by incorporating multidimensional endpoints from various sources (e.g., registries, claims, EMR, epidemiology)
- **HTA bodies expect triangulation and robustness,** are scrutinizing data lineage and robustness, and often disdain single source datasets
- **Support for rare disease/subset patients** usually require stitching multiple fragmented information into a cohesive dataset

If you're in Clinical Development:

- **Define optimal target populations for recruitment and enrollment** based on more robust patient estimates, especially in smaller or more targeted disease states and countries/regions

- **Identify better trial sites** using global epidemiology overlays
- **Design smarter inclusion criteria** by understanding real-world heterogeneity
- **Accelerate recruitment** by targeting true prevalence clusters.

Final Thoughts: Moving From Data Scarcity to Data Strategy

In cancer and rare disease research, data is never perfect. But the smartest teams know how to make the imperfect work. That means pulling together multiple perspectives, cross-checking assumptions, and building estimates you can actually trust.

Using **multiple sophisticated epidemiology sources** isn't just a nice-to-have. It's **a strategic imperative** by providing practical, risk-mitigating, and decision-enhancing insights. Especially in high-stakes therapeutic areas where every patient counts, every signal matters, and every insight could shape the next breakthrough.

The ability to take rigorous historical epidemiology data and combine it with estimated future shifts in incidence, staging, survival, and treatment assumptions to **generate future "scenarios"** of patient numbers could provide you with a profound competitive advantage.

So, the next time someone tells you they have "the" patient number for your target indication. Ask them: "Based on what sources?" Because the answer could make all the difference.

Need help identifying the right epidemiology sources or assistance in implementing the integration/triangulation process for your next asset strategy? Reach out to the epidemiology experts at Epiphany Partners (https://time4epi.com, Brittany Britt, Business Development, brittany@time4epi.com)

Citations

Lawlor DA, Tilling K, Smith DG; "Triangulation in aetiological epidemiology;" Int J Epidemiol, 45(6): 1866-1886, Jan 2017.