graticule™

White Paper

THE USE OF RWD TO ASSESS SAFETY OUTCOMES IN THE US AND EU

graticule.life

Introduction

As drugs reach the US and EU markets at record pace, driven in part by AI-powered drug discovery and advances in monoclonal antibody design, the number (and complexity) of safety studies is growing in parallel[1]. Regulators in both regions, increasingly aware of the availability of real-world data from electronic medical records, are expecting more robust safety surveillance powered by access to granular clinical data. Pharmaceutical companies thus need scalable, sustainable approaches to safety monitoring that meet regulatory expectations for timeliness, completeness, and geographic coverage.

The secondary use of patient medical records offers a promising path forward. Real World Data (RWD) refers to the analysis of data originally collected for purposes other than research, such as billing, clinical care, or public health monitoring to generate real-world evidence (RWE) about the safety and effectiveness of medical products. Such RWE captures clinical outcomes as they unfold in routine practice, without the intensive oversight, monitoring, and protocol-driven controls imposed in clinical trials. Consider complex inhalers or multi-dose-per-day drug regimens: effective under supervision; error-prone in the real world. Moreover, the statistical power needed to detect meaningful safety or effectiveness signals may only be achievable through large-scale data repositories like national claims databases or health system electronic health records (EHRs) spanning many delivery networks.

In addition to the unique insights it can provide, the secondary use of patient data can be accessed faster and more cost-effectively than its prospective data counterpart, which is muddled by logistical and regulatory hurdles associated with informed consent and patient management (scheduling, follow-up, interviews, questionnaires, etc.).[2]

This efficiency is particularly valuable to life science companies in the context of FDA and EMA safety reporting requirements, which emphasize timeliness and broad demographic coverage[3]. Reflecting this need, both agencies are actively embracing RWD for regulatory use. The FDA has a longstanding history of leveraging real-world data to monitor and evaluate postmarket drug safety[4]. Similarly, the EMA has established the Data Analysis and Real World Interrogation Network (DARWIN EU), operationalized in 2024, to support regulatory decision-making through the use of

---

[1] https://aspe.hhs.gov/number-us-fda-anda-approvals-fiscal-year
[2] https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-022-01768-6
[3] https://www.ema.europa.edu/
[4] https://www.fda.gov/drugs/cder-conversations/real-world-evidence-safety-potential-tool-advancing-innovative-ways-develop-new-medical-therapies/

high-quality, real-world data sources across Europe[5]. Medical devices in particular have increased requirements including the MDR (Medical Device Reporting) enacted in 2017 in the EU adding a need to provide relevant safety data from European patients for devices to maintain regulatory authorization.

While there are benefits to using RWD for clinical purposes, it is nevertheless subject to numerous biases. Real-world data sources are not all created equal, and the choice of dataset or approach to access can fundamentally shape the quality, interpretability, and regulatory acceptability of a study. Claims and EHR data have distinct advantages, but also come with trade-offs in terms of clinical detail, completeness, longitudinal follow-up, and data governance (particularly in the EU) - not to mention cost and data access timelines that vary by geography or data access approach.

In the sections that follow, we outline the strengths and limitations of these data types in the US and EU, with a focus on their suitability for safety studies in an evolving evidence landscape.

[5] https://darwin-eu.org/

Overview of Secondary Data Sources

Claims Data

Administrative claims databases are structured collections of billing and reimbursement records generated through insurance coverage. In both the US and EU (especially in countries with single-payer systems), claims datasets can provide large-scale, population-level insights into medication exposures, healthcare utilization, and coded clinical outcomes.

Claims can be further subdivided into private and public payors, though this distinction is more relevant in the US where income, age, and job-status largely determines payor status. For instance, individuals over 65 are typically covered by Medicare, a federal public insurance program, while employed individuals often receive employer-sponsored commercial insurance. This is in contrast with the EU-5, where over 90% of the population is enrolled in public insurance.[6]

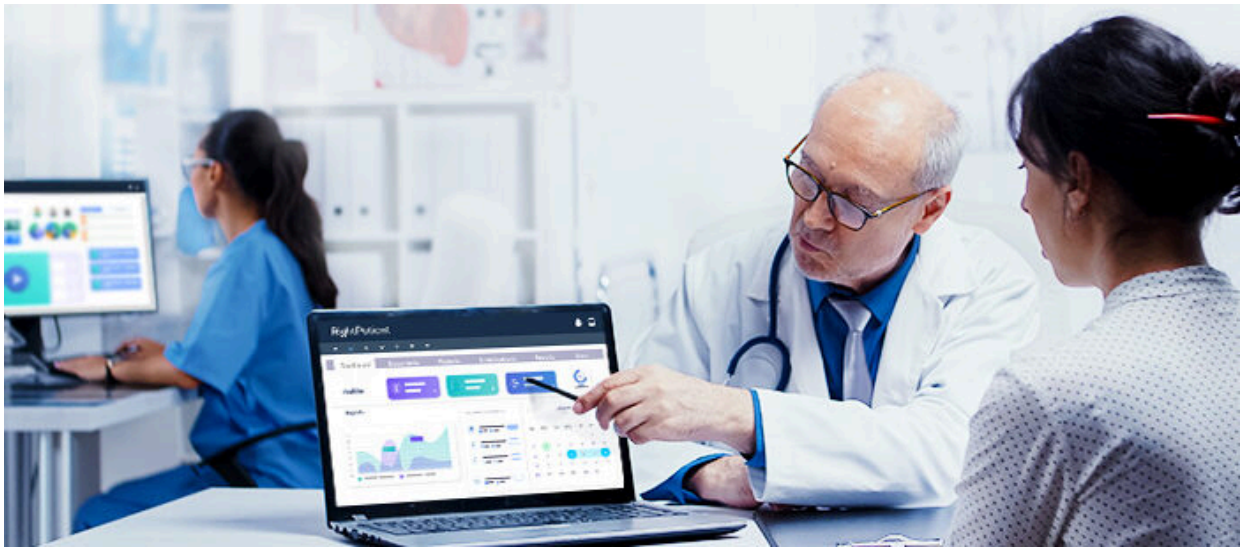
Historically, many RWE safety studies have used claims or related administrative data. The advantage of claims is often the scale of coverage of patients and longitudinal information about them despite changing provider settings. However, claims data may

[6] https://pmc.ncbi.nlm.nih.gov/articles/PMC8005513/

have limitations for the objectives of safety studies. The primary challenge is that both identification of patient populations for studies and safety endpoints are not available in claims. While a readmission or an effect such as ICU or ED visit will be available in claims, they do not collect any patient related complaints that can't be encoded in a structured procedure, diagnosis, or prescription. They often provide only a coarse view regarding the reason for the visit such as admission diagnosis that can be difficult to directly correlate with product use.

Adverse events of interest for safety are most often documented in standardized instruments by clinicians such as scales for pain scales, blood pressure measurements, weight gain, and radiology reports. Patient reported AEs such as headache, GI distress, nausea, loss of breath, or cognitive deficits may only be visible in free text notes or patient reported outcomes. Lab results or observations which demonstrate dysregulation such as elevated markers of kidney require not just that a test was completed but also the specific test scores, measurement calibration to standards, and changes from baseline values for the patient in order to interpret the effects in the context of safety. Overall as a result claims data may be found insufficient for achieving a goal to detect a key adverse event or signal relating to known risks to a medical product.

Electronic Health Records (EHRs)



EHR data are derived from routine clinical care and include patient demographic and clinical characteristics, treatments and procedures, physician notes, lab data, and imaging reports. These data offer rich clinical context, especially useful in complex safety assessments where lab results, biometric values, clinical notes, and other

parameters are necessary to derive outcomes. Like claims data, EHR data access is mediated through two primary channels: EHR data licensing and direct site access.

Commercially-available datasets represent the most expensive and rapidly accessible EHR data option. In the US, commercial RWD vendors have successfully produced subscription products that aggregate anonymized data from multiple health systems or provider groups into unified datasets. These offerings vary widely on the scale and pricing models. Vendors often differentiate themselves by focusing on specific therapeutic areas, targeting niche care settings (psychiatry, oncology, ophthalmology, community practices), integrating additional modalities (e.g., lab results, imaging, specimen genetic sequencing), or enriching datasets with AI-derived variables. Flatiron Health's oncology and hematology datasets in the US is a prominent example of a licensable data set in the US. It focuses on data collected from community oncology using the FlatIron EHR with significant enrichment for oncology concepts. Some vendors emphasize structured vs. unstructured data capabilities, or curated outcomes for specific diseases by enriching data by building pipelines to map text into distinct fields or conducting manual abstraction by experts from notes. Formats for these data sets are often proprietary but are at times normalized into common data models such as OMOP.

In contrast, commercially-available EHR data in the EU are more limited, with only a few vendors offering datasets focused on narrow therapeutic areas. The regulatory framework of GDPR and fragmentation of European geographies have made it difficult to generate licensed data. In some geographies such as the UK (CPRD and UK Biobank) and Finland (Findata) licensable data has been established through government programs. But similar capabilities are not available in key markets such as France, Germany, Spain, and Italy. Much of the EHR data infrastructure in Europe remains decentralized or government-driven as public projects or public private partnerships such as the IHI programs including EHDEN.

Direct site access (whether in the U.S. or EU) refers to establishing partnerships with individual hospitals or integrated delivery networks (IDNs) to extract and deidentify the source EHR data. This approach offers the highest fidelity and flexibility, especially for unstructured content like clinician notes, scanned documents, or detailed prescribing patterns. The increased flexibility comes from working closely with health system IT groups to establish the specific access to the data resources needed for each research project. Furthermore by using data use agreements rather than licenses the use of data can extend for longer periods for multi-year studies. Costs billed from health systems are generally related to actual costs vs. a profit model from the licensing. These approaches can significantly lower the long term cost for data use. However, direct site access often requires additional governance, infrastructure, and coordination that may

delay the study start at each site or the overall timeline from coordination across many sites to achieve sufficient scale to support the safety goals.

In the U.S., access through direct partnerships is typically managed through establishment of research protocols reviewed by the health system IRB and involves collaboration with site IT and clinical teams including a principal investigator at the site. In the EU, site access is more fragmented and highly country-specific. Some countries, like the Netherlands and Finland, have made progress through national or regional integration efforts. For instance, Finland makes available in a single access hub country-wide registers of claims, EHR, and social-demographic datasets with nearly complete national coverage. Other countries rely on federated models, such as Germany's Medical Informatics Initiative or France's Health Data Hub. Access generally requires site-by-site engagement, approvals from ethics committees or national data custodians, and full compliance with GDPR. While pan-European EHR platforms are beginning to emerge, much of today's research still depends on federated analytics or direct partnerships with national health data organizations.

Table 1: A comparison of claims and EHR data types

Data Source	Subtype	Region	Scale	Standard-ization	Clinical Depth	Lag Time	Time to Access	Cost of Access
Claims	Public	US	National	Yes	Low	6 Months	1 Week	$$$
		EU	National	Yes	Low	12 Months	6 Months	$
	Private	US	National	Yes	Low	12 Months	1 week	$$
		EU	National	Yes	Low	12 Months	6 Months	$
EHR	Commercial	US	National	Yes	High	~1 Week	1 Week	$$$$
		EU	N/A[7]					
	Direct Site Access	US	Health-System Level	No	Highest	~24 hours	3-6 Months	$$
		EU	Site-Level	No	Highest	~24 hours	6-12 Months	$$

Scale refers to the maximum possible coverage achieved in a single dataset (or single site data warehouse integration). *Standardization* refers to data that has been transformed to follow a common data model (CDM) or uniform format—allowing for

[7] US data firms like Flatiron have expanded into Europe, but currently serve narrow therapeutic areas and have similar cost and timeline structures as their US counterparts

consistent analysis across different systems or sources. Clinical depth refers to inclusion of unstructured data elements, lab results, and other test outcomes. Lag time describes the frequency of data updates.

Figure 1a: Cost vs. Time-to-Data-Access of US and EU RWD Datasets for secondary use of data

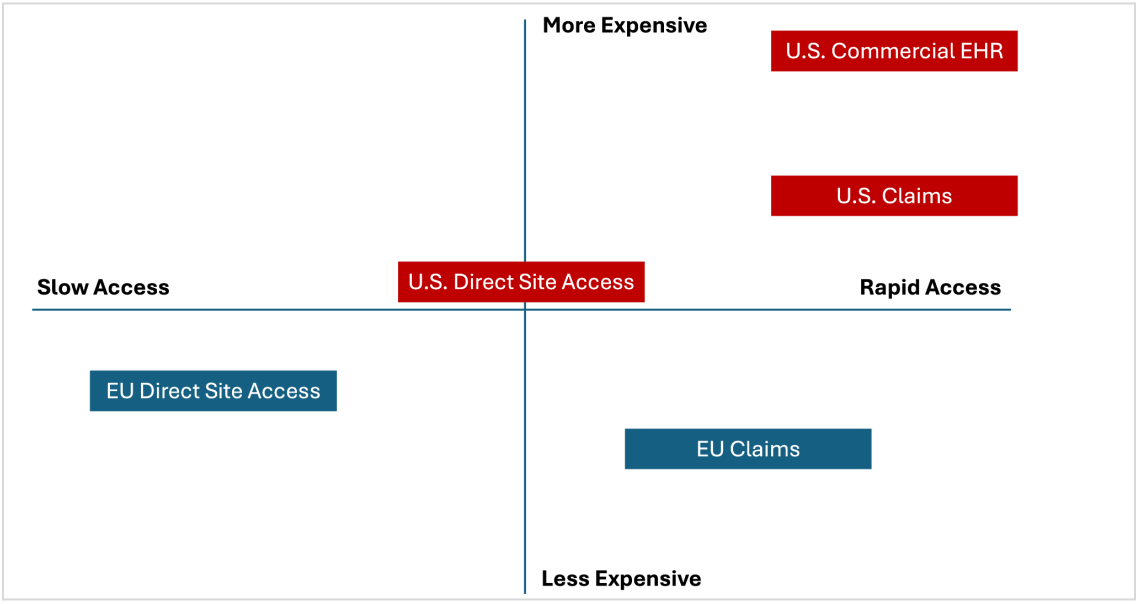
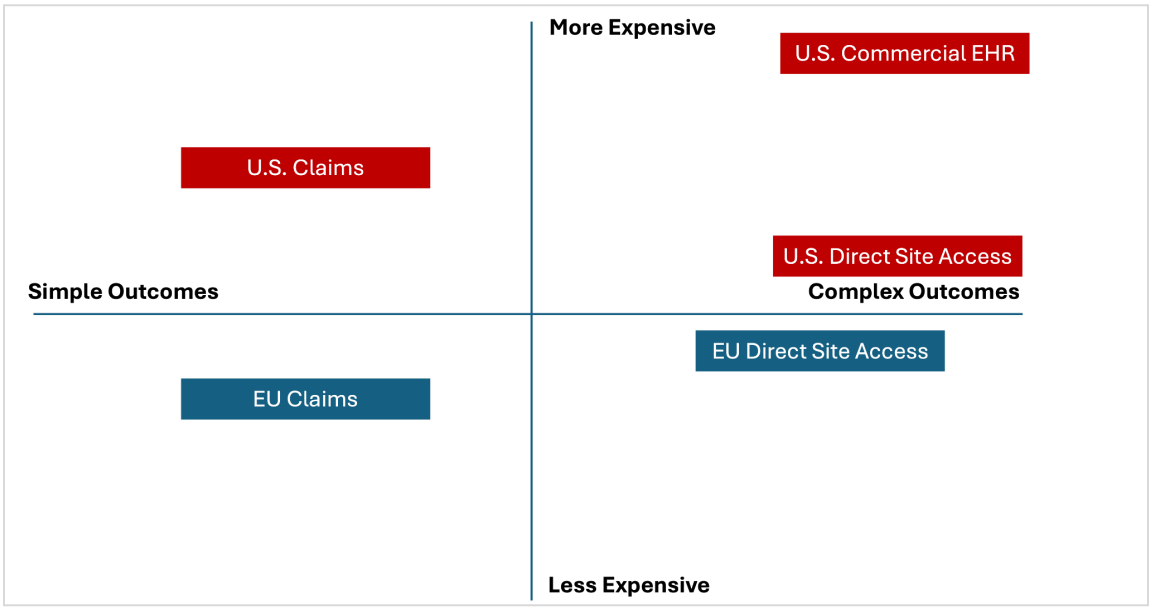


Figure 1b: Cost vs. Outcome Complexity Analysis of US and EU RWD Datasets for secondary use of data



Identifying Fit-for-Purpose Data for Safety Studies

Given the breadth of data options, finding the 'right' data can be daunting. EMA requires data collection for up to eight years when imposing a post-authorisation safety study (PASS), highlighting the importance of selecting data sources that are reliable, sustainable, and fit-for-purpose over the long term.

Naturally, this process of identifying the right data source depends on the research question and outcomes of interest, and is further influenced by budget, timeline, and resource constraints. We repeatedly find that evidence teams assume or default to known or previously used data without fully considering whether that data can support the nuance of the research question, in turn forgoing a de-risking data landscaping process. This can lead to costly missteps, including inconclusive results or the inability to satisfy regulatory requirements.

While recent examples of this are not yet in the public domain, the FDA and EMA have been increasing scrutiny on real-world evidence submissions that lack adequate justification for data source selection. The FDA's 2023 draft guidance emphasizes that sponsors must assess whether a RWD source contains "sufficient detail to capture the information needed to evaluate the question being addressed in the target population" and cautions against proceeding without confirming the data's completeness and reliability.[8] Likewise, EMA's GVP Module VIII notes that the appropriateness of data sources should be evaluated based on validity, completeness, and alignment with study objectives, urging sponsors to justify the choice of dataset in relation to the study question.[9]

We thus recommend a structured framework for data landscaping. Our approach begins by clearly defining the research objectives and outcomes of interest. We then assess the feasibility of answering the question starting with lower-complexity sources such as claims data. As discussed, these sources are often sufficient for tracking diagnoses, medication use, or healthcare utilization. However, for many safety outcomes, like those requiring lab-confirmed diagnoses (e.g., C. difficile infections in antibiotic safety studies), claims data will fall short. In such cases, it becomes immediately apparent that either 1) granular EHR data is required to reliably identify the outcome or 2) the outcomes of interest must be revised.

[8] https://www.fda.gov/media/174819/
[9] https://www.ema.europa.eu/en/homepage

To avoid these missteps, we employ a stepwise data selection algorithm that begins with the simplest viable source and escalates in complexity only as needed. This approach balances efficiency with precision and helps teams justify their data strategy in line with evolving regulatory expectations.

Stepwise Algorithm for Selecting Fit-for-Purpose Data

1. **Start with the research objective**
    - What is the safety signal or outcome of interest?
    - Is it a hard endpoint (e.g., mortality, hospitalization) or a soft signal (e.g., adverse event symptoms, lab abnormalities)?
    - Is the population common (e.g., T2DM patients) or rare (e.g., CIDP, pediatric SLE)?

Rationale: A clearly defined question anchors the entire data selection process. Safety studies vary widely in required granularity: detecting mortality or hospitalizations may require simple coded data, while detecting subclinical adverse events or rare complications may demand richer clinical inputs. Clarifying population, exposure, outcome(s), and covariates up front prevents downstream misalignment.

2. **Can claims data sufficiently capture the required exposure, population, and outcome?**
    - **Yes** → Proceed with feasibility and completeness assessment of relevant claims datasets.
      No → Move to EHR-based solutions.

Rationale: Claims datasets offer fast, scalable access to structured, longitudinal patient data and are thus ideal for common outcomes (e.g., ED visits, surgeries, medication fills). In comparison to EHR, claims data minimize costs and time-to-access. However, they lack clinical detail (e.g., lab outcomes, clinical notes) and are often updated at delayed intervals (e.g., data refreshes quarterly or semi-annually).

3. **Is structured EHR data (from commercial aggregators) sufficient?**
    - Check whether structured fields (diagnoses, labs, procedures) cover the exposure and outcomes.
        - **Yes** → Use commercial EHR dataset (e.g., Flatiron, TriNetX, Cegedim) and validate with feasibility queries.
        - **No (or if Cost-Prohibitive)** → Escalate to direct-site EHR access for richer clinical context.

Rationale: Structured EHR data adds clinical richness (labs, vitals, genetic outcomes, etc.). These datasets can be accessed more quickly than direct-site partnerships and often suffice for studies involving moderate clinical detail. However, commercial datasets are expensive and typically operate on licensing models - a particular concern for 8-10 years PASS where data fees alone can exceed 5 Million USD. Note that commercial data sets span academic networks, community hospitals, and also specialty practices. Identifying the 'right' commercial EHR data can be challenging for ambulatory care.

4. **Is direct site access needed?**

    - Identify the type of sites most likely to record the required data:
        - **Academic hospitals** for rare diseases, complex diagnostics, or high-resolution longitudinal follow-up
        - **Specialty clinics** (e.g., neurology, oncology) for subspecialty care and domain-specific assessments
        - **Integrated delivery networks (IDNs)** for continuity across inpatient and outpatient care
        - **National registries** or **EHR-claims linkages** if available in relevant geography (e.g., Sweden's NPR or France's SNDS–EHR linkages)

Rationale: Direct-site access offers similar data to commercial datasets but operates on a different cost model - typically site resource and effort vs. licensing. For this reason, direct data is often cheaper to access and use for the duration of a study (this is critical for safety studies where patients must be tracked longitudinally). However, access is governed via IRB and the accessor is often responsible for data cleaning and harmonization - increasing the burden of work on the study team.

5. **Assess data governance and sustainability**

    - Will the source permit reuse across years for long-term PASS obligations?
    - Are there barriers related to cost, contracting, ethics approval, or GDPR?
    - Is it possible to extend access if regulators request supplementary analysis or re-submissions?

Rationale: This step evaluates whether the data source permits extended access, data reuse, and reanalysis critical for meeting EMA/FDA requirements. Sponsors must also ensure data access aligns with privacy regulations (e.g., GDPR) and includes rights to fulfill future regulator requests.

The Current Status of Safety Studies in the EU

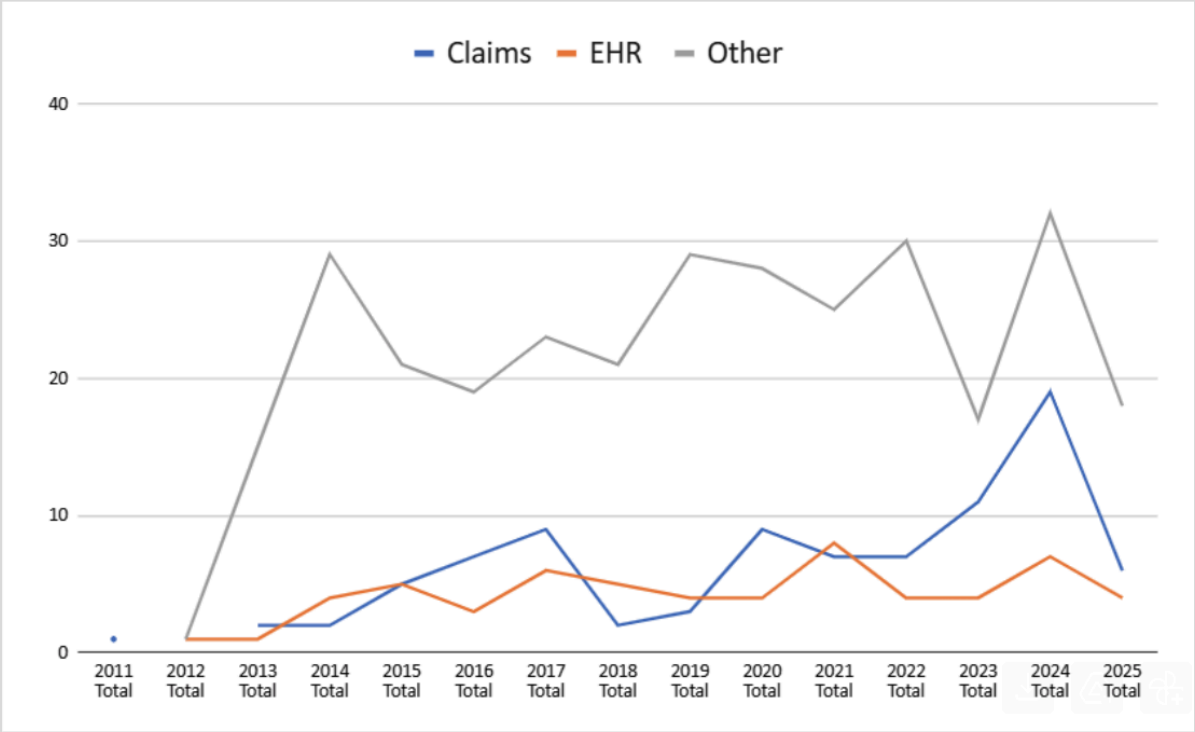
Given the expansion of novel drugs and biologics, particularly in immunology and oncology, pharmaceutical developers are increasingly seeking not only regulatory approval but also product differentiation in terms of performance (e.g., comparative effectiveness) and safety (e.g., adverse event profiles across subpopulations). Consider the IL-23 inhibitor class: as of 2024, multiple agents such as Guselkumab, Risankizumab, and Tildrakizumab are approved for plaque psoriasis and/or psoriatic arthritis, with ongoing head-to-head trials and indirect comparisons aimed at teasing out nuanced differences in durability, onset of action, and long-term safety[10]. As these agents compete within crowded therapeutic landscapes, the demand for real-world evidence that can support fine-grained comparisons continues to grow.

One would expect a proportional increase in the use of EHR data given its capacity to capture detailed clinical outcomes, lab values, and adverse events documented in free-text notes or imaging reports. However, this has not occurred in Europe. As shown

[10] Armstrong AW, Read C. Pathophysiology, clinical presentation, and treatment of psoriasis: a review. JAMA. 2020;323(19):1945–60. doi:10.1001/jama.2020.4006.

in Figure 3, claims and registry data have continued to dominate EMA-mandated PASS studies from 2012 to present. Rather than reflecting the unsuitability of EHR data, we interpret this pattern as a byproduct of persistent access barriers, limited harmonization, and the comfort of study teams with long-established registry and claims infrastructures.

Figure 3: Proportion of PASS Studies in the EU utilization Claims, EHR, or Other Data sources



Despite EMA pressure to provide nuanced outcomes for PASS and other safety analyses, claims and registry (among other data types) represent the majority of EU PAS studies. As discussed, EU initiatives such as DARWIN EU, the EHDEN Foundation, and national data integration programs (e.g., Findata in Finland, Health Data Hub in France) are laying the groundwork for secure, scalable, and governance-compliant access to EHR data across member states. The maturing of these platforms combined with increasing regulator openness to EHR-derived evidence will likely catalyze a sharp increase in their use for safety studies. As the technical and administrative hurdles diminish, and as research teams gain familiarity with federated analytics and data standardization frameworks (e.g., OMOP CDM), we expect EHR data to shift from niche to norm, particularly for studies requiring deep clinical phenotyping or real-time pharmacovigilance.

Recommendations and Considerations

Regardless of geography, early initiation of data access planning is essential, and especially critical in Europe. The decentralized nature of governance, country-specific ethics processes, and evolving data infrastructures mean that timelines can stretch significantly if not anticipated from the outset. It is not uncommon to receive data after 8-12 months of contracting. Starting conversations with data partners and ethics committees during protocol development can save months downstream.

Timely and cost-effective data use can be further supported via a hybrid data access strategy, when possible. Sponsors should consider licensing commercial EHR datasets early to run feasibility queries, test codelists, and develop preliminary algorithms while concurrently initiating direct access to site- or registry-level data. This parallel approach ensures that by the time the richer but slower-to-access site data is available, the study team has already validated endpoints, cleaned exposure definitions, and prepared (to some extent) analysis-ready code. This de-risks study execution and reduces the chance of protocol amendments due to data structure mismatches or feasibility issues.

Similarly, consider the long-term reusability of the dataset as another de-risking measure. EMA-mandated PASS studies often require follow-up across 5–8 years, during which time regulators may request interim analyses, subgroup assessments, or post-hoc evaluations. Sponsors should ensure that their chosen data partners can support re-analysis, maintain access continuity, and provide ongoing data refreshes throughout the lifecycle of the product. They should also plan for data changes during this time - EHR migration, ontology switches (e.g., impending ICD-11 uptake in 2027), and hospital mergers.

Finally, it is critical to engage multidisciplinary teams early—including data scientists, epidemiologists, legal and compliance experts, and local country affiliates—to align on the feasibility, cost, and scientific acceptability of the proposed data sources. In many cases, small investments in upfront planning, simulation modeling, or pre-alignment with regulatory authorities can prevent far costlier mid-study redesigns or non-acceptance of study findings.

Interested in learning more?
Reach out to us at info@graticule.life