Verana Health®

Strengths and Limitations of Real-world Data from Electronic Health Records and Claims, and the Value of Combining Data from These Sources to Inform Research

The broad adoption of electronic health record (EHR) systems and advances in data curation techniques are providing researchers another valuable observational data source beyond the long-relied upon medical and pharmacy claims data to answer important questions in the health care ecosystem. Life sciences companies have utilized claims data to deliver insights into patient populations and treatment utilization in real-world practice to guide the development, launch, and commercialization of new therapeutics and medical devices. However, due to the inherent nature of what is captured through claims, combined with a rapidly expanding understanding of the biologic and genetic underpinning of various diseases, this datasource, used alone, is not always fit for use. Claims data, while able to capture care utilization across the health care system, may lack key clinical characteristics, outcomes of interventions, patient experience, results of diagnostic evaluations and physician decision-making. There is an opportunity in many diseases to enhance insight capability by linking EHR and claims databases. These two different, but complementary datasets can help to achieve a more complete picture of disease trajectory and management.

There is an opportunity in many diseases to enhance insight capability by linking EHR and claims databases. These two different, but complementary datasets can help to achieve a more complete picture of disease trajectory and management.

One of the main uses of a combined EHR and claims database is to enhance the depth of the available data model. Oncology is one therapeutic area where the availability of large EHR-derived datasets has facilitated research and advanced the positioning of real-world evidence. Using these novel datasets, researchers can add non-codable disease characteristics such as a grade, histology, or stage of a cancer, which are critical to understanding cancer biology and treatment options. Linking EHR data with claims data in this context provides valuable information about multidisciplinary treatment received across institutions or providers. Stated differently, combining EHR and claims datasets could provide a solution to known limitations in each. For example, claims pharmacy data can be used to investigate concomitant medications that are not always accurately recorded by a single provider or found in a single EHR system, given that patients are often prescribed medications by multiple physicians. Another use case is combining the detailed documentation on patient symptoms, results of diagnostic evaluations, and physical exam findings within the EHR with therapeutic or device utilization data from claims to investigate more complex outcomes questions.

For life sciences companies, these clinical examples translate into deeper insights across all stages of the drug development lifecycle, including clinical development, market opportunity, trial planning, post-approval, and commercialization. This powerful linked EHR and claims dataset can help life sciences companies identify subsets of patients with limited treatment options, evaluate market performance and safety, and understand physician-prescribing and post-approval treatment patterns across different therapeutic areas.

This powerful linked EHR and claims dataset can help life sciences companies identify subsets of patients with limited treatment options, evaluate market performance and safety, and understand physician-prescribing and post-approval treatment patterns across different therapeutic areas.

When using secondary observational data, it is critical to understand strengths and limitations of any dataset, including why the information is documented and then decide how to optimize fit-for-use in databases. A key consideration is to decide when the linked dataset is more reliable and complete in the context of a clinical or research question as opposed to relying on either independently. The linked datasets can create a deeper data model, but will come with tradeoffs that need to be considered. This paper will review some basic concepts around each data source and examine the capabilities and limitations of a linked EHR and claims database versus data derived from these individual sources. Consideration for the optimal data source is both use-case and disease dependent. Figure 1 is an example of a framework for determining the most suitable data source(s) by use case and disease.

Figure 1

Example of research use cases in open-angle glaucoma with an implantable MIGS device, and which RWE data sources may be best suited*

Example Use Case	EHR	Claims	EHR + Claims
Trial Design & Optimization	✓	✓	✓
Clinical Trial Site Selection	✓		
Clinical Trial Patient Identification	✓		
Health Resource Utilization & Modeling		✓	
Natural History of Disease	✓		
Postmarketing Safety Study	✓		
Burden of Disease	✓		
Treatment Patterns & Outcomes	✓	✓	✓
Comparative Effectiveness	✓		
Drug Utilization		✓	
Market Share Tracking	✓	✓	✓
Label Expansion & Modification	✓	✓	✓

*All use cases require feasibility on a case-by-case basis and may not be representative of all research project requests.

Claims Data

Definitions

- **"Open" (or "non-payer complete")** claims data can be sourced from any number of non-payer institutions, including physician sources (e.g., practice management systems, EHRs), pharmacy benefit managers (PBMs), and clearinghouses (electronic stations or hubs that allow healthcare practices to transmit electronic claims to insurance carriers).

- **"Closed" (or "payer-complete")** claims data are derived from an insurance provider (a.k.a. "payer"), and captures nearly all the procedure and treatment events that occur during the patient's enrollment period (e.g., medical and pharmacy visits and transactions).
    - Continuous enrollment: time window around the index date during which a patient will be considered closed or payer-complete.

Background

There are two sources of claims data routinely available for research.[1] The first is **medical claims data**, also known as administrative data, which are collections of information on millions of doctors' appointments, bills, insurance information, and other patient-provider communications. Medical claims data contains specific medical codes that detail the care administered by providers during a patient visit. These codes typically include procedures, diagnoses, medical devices, and prescription drugs.

The vast number of patients included in commercially available claims databases combined with the structured nature of the available data creates an excellent opportunity for research. As a result, claims data fundamentally represents patients' interactions with the medical system that are submitted for payment, and delivers a comprehensive economic view because it represents costs in the healthcare system. Claims data, such as International Classification of Diseases (ICD) and Current Procedural Terminology (CPT®) codes, work well at capturing structured information about procedures, drug delivery, and general utilization of resources due to standardization and oversight from various governing bodies.

Information about prescription drugs are also found in **pharmacy claims.** Pharmacy claims are generally those drugs that are self-administered and include most prescription oral and self-injectable drugs. Specialty pharmacy medications can be found in different sources depending on the insurance plan and this needs to be understood if relevant to the research question. By bringing together different elements found in claims databases, a patients' clinical journey can be further investigated. For example, in a patient with a degenerative neurological disease such as Parkinson's disease, claims data can provide a longitudinal view on medications received over time, utilization of diagnostic tests, potential hospitalization, and use of durable medical devices such as a wheelchair. The latter could provide an important surrogate for progression of disease.

[1] https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence

> **Claims data fundamentally represents patients' interactions with the medical system that are submitted for payment, and delivers a comprehensive economic view because it represents costs in the healthcare system.**

Potential Limitations

Despite representing large populations of patients, structured codes that are used in claims lack the level of granularity needed to understand important clinical characteristics such as disease subtypes, nuances of disease progression, symptoms that are not coded, reasons for treatment discontinuation, and clinician decision-making.

Indeed, the granularity of claims data is only as fine as what diagnostic and procedure codes enable and exactly how clinicians use those codes to facilitate reimbursement. For example, if multiple diseases or brands of medical devices are rolled up into a single ICD-10 code, claims data cannot be used to distinguish between them. An example is solid tumor cancers where the ICD-10 code reflects the anatomical location of the tumor but not the stage, histology, or mutation profile of the tumor— relevant for many research questions. Another example is genetically defined diseases such as spinal muscular atrophy (SMA) where the population is now better sub-divided by underlying genetic subtype, a key disease characteristic which does not map to individual diagnostic codes.

While claims data does indicate when a provider ordered a certain diagnostic test for a patient, it does not record what triggered the ordering of the test or its results. Additionally, claims data may include erroneous diagnostic information as a result of coding that was performed to obtain reimbursement, but does not reflect the actual diagnosis.

Overall, claims data often lacks valuable clinical nuance, so much of working with claims data involves looking for patterns and making assumptions to create rules around those patterns. For example, if a patient stopped or switched a drug, then a researcher would have to make assumptions on the reasons for discontinuation which could include development of an adverse reaction to the drug. Other reasons, however, could include financial burden to the patient, lack of efficacy, patient refusal, or interaction with other medications.

Similarly, claims data is not useful in capturing the broad range of potential drug toxicity. Claims data will provide documented symptom codes, prescriptions, hospitalizations, or a procedure to address the toxicity. However, adverse reactions that are managed over the phone or with over the counter medications may not be as well captured in billing codes and will be unknown to the researcher. In addition, claims data cannot help researchers understand how much of a drug was taken, attempts at titration, etc. Simply, claims data indicates that a prescription for a drug was filled and then later not refilled.

Finally, there are administrative considerations around commercially available claims data that must be understood. To be included in a database of this nature, a patient has to have insurance, excluding uninsured patients. Commercially available claims databases aggregate as much information as possible from any available source to provide the largest available datasets. This results in both open and closed claims databases. An additional note is that we can also think of an individual patient as being "open" or "closed" for a particular period of time. For example, a patient would be considered "closed" from X date to Y date because one is able to obtain claims from a payer with whom they were enrolled for that period of time. Studies often require that patients be closed for a certain window around an index date. Closed claims, generally considered more reliable to represent what health utilization actually occurred, have some important gaps in data availability. For example, the only claims available on patients with Medicare will be those with Medicare Advantage, missing the basic Medicare claims information. Second, there is generally no data from the Veteran Affairs included in claims databases. This bias that these nuances introduce into analysis should be considered when selecting fit-for-use datasets and outlining the limitations of the generalizability of the results.

[2] https://openpublichealthjournal.com/contents/volumes/V2/TOPHJ-2-11/TOPHJ-2-11.pdf

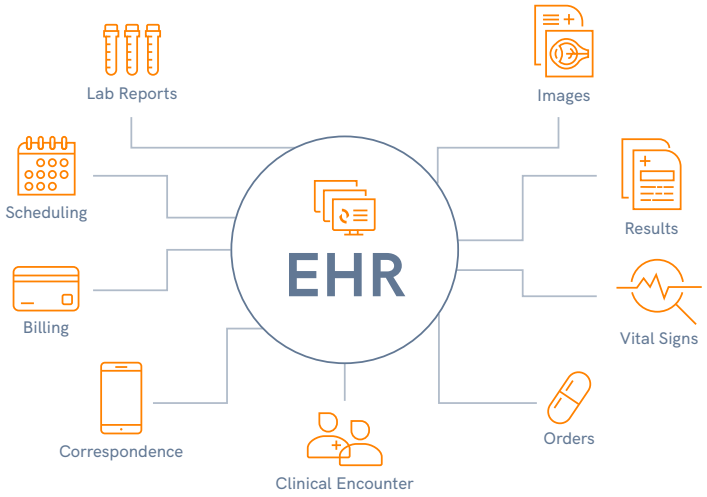
Electronic Health Record (EHR) Data

Definitions

- **Structured data:** information contained in drop-down boxes and specific fields such as "current medications," "date of birth," or "race/ethnicity." These are often standardized fields across EHR systems, such as within a certain medical specialty.
- **Unstructured data:** generally free-form text information entered by clinicians that appears in the notes sections of EHRs that range from symptoms to outcomes, such as "ambulatory status," "response to therapy," and "progression of disease."

Background

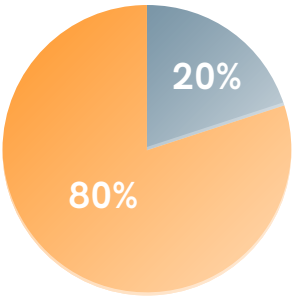
To obtain a more comprehensive understanding of disease, including physician documentation of a patient's symptoms and management planning, one could utilize EHR data. EHR data contains a patient's schedule, medical history, diagnoses, medications, treatment plans, immunization dates, allergies, radiology images, and laboratory and test results.



All data in an EHR is entered in either a structured or an unstructured field. Structured data generally reflects data that is entered into pre-specified fields or selections such as check boxes or drop-down choices. It also reflects information that is pulled into an EHR, such as order entry, laboratory values, and scheduling information. While there is some work needed to clean and harmonize structured data for it to be research ready, the transformations are generally standardized across various EHR systems and are generally conducted using an accepted ontology. There has been a push for more medical data to be entered into the EHR in a structured format for ease of processing and analysis, but adoption has been slow for a multitude of reasons, most notably physician reluctance related to its negative impact on their productivity.[2]

Therefore, for the foreseeable future, most of the information in an EHR will be found in an unstructured format. Unstructured data is any information that is entered into a free text format, and it is estimated that about 80% of medical data is unstructured. This importantly includes clinical notes where physicians routinely document sentiment behind the actions – specifically, what that patient is feeling and what the clinician is thinking. It can also include information not accounted for in structured fields, such as genetic testing results, reasons for changing medications, and external factors that may impact the health of a patient. This type of information could be very valuable to study side effects or tolerability of medications, dosing modifications, and potentially less severe complications of a procedure not necessarily captured in claims data.

80% of medical data remains unstructured*



[3] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5510605/
*https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6372467/

Data that appears in the notes section can provide deeper insight into many additional aspects of a patient's disease journey, such as what happened to the patient prior to a diagnosis or why a clinician chose a certain therapy or medication. It can also offer relevant details about a condition where specific coding may be lacking, such as smoking or undiagnosed conditions like hypertension, which would be captured by evaluating blood pressure readings over time. Additional details from within other types of notes are also available, such as the type of device used in a certain procedure from an operative note or genetic alterations that might be driving a disease from a sendout laboratory report.

Potential Limitations

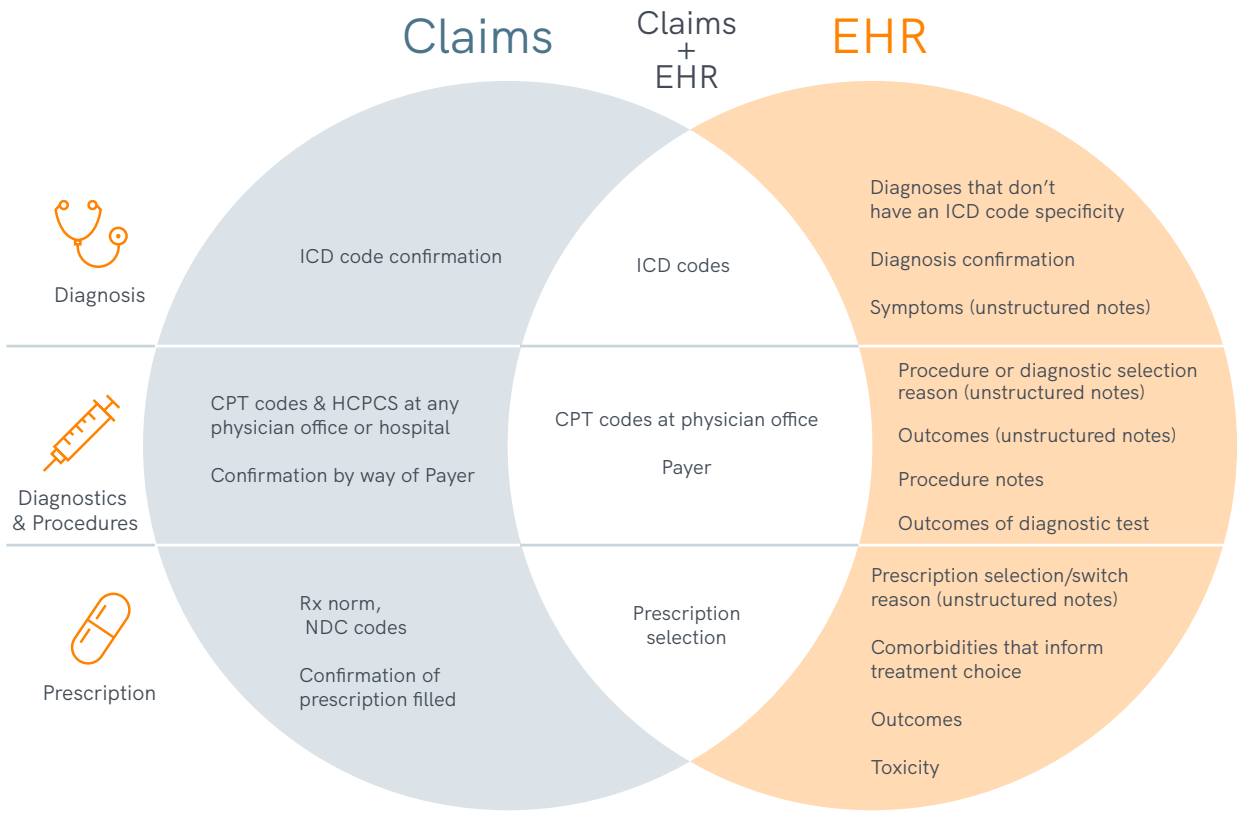
While EHR data provides a wealth of valuable information, it also comes with some inherent limitations. When something is not documented, it does not mean that it never happened; it simply means it was not included in the record. Consequently, the data available in EHRs is only as good as the level of documentation by the clinicians.

For example, EHR systems are often tied to provider specialty, such as ophthalmology, urology, and neurology. As a result, some EHRs deliver a significant amount of granular information about what a patient is experiencing within the context of that particular provider – such as eye issues for ophthalmology patients – but may lack information on the rest of the patient that exists outside the context of that one provider's specialty, such as symptoms or medications related to their comorbidities like diabetes.

Additionally, most EHR systems focus on servicing outpatient physician practices because this is where the majority of care in the U.S. occurs today, so some do not capture patient data related to hospitalizations, physical therapy or other allied support, which are an important endpoint for many clinical studies.

Lastly, using EHR data to inform real-world research can come with delays or limitations due to data quality. To make EHR data research-ready— especially bringing structure to the vast majority of it that is unstructured—requires deep expertise and process around data cleansing, harmonization, and curation. It also calls for deep clinical knowledge to interpret text-based documentation and experience with artificial intelligence (AI) techniques, such as machine learning (ML) and natural language processing (NLP), to standardize this data in a meaningful way at scale.

Data availability by source



ICD Code: International Classification of Diseases (ICD) is a globally used diagnostic tool for epidemiology, health management and clinical purposes managed by the World Health Organization.

CPT Code & HCPCS : Healthcare Common Procedure Coding System (HCPCS) is a set of health care procedure codes based on the American Medical Association's Current Procedural Terminology (CPT).

NDC Code: National Drug Code is a unique 10-digit or 11-digit, 3-segment number, and a universal product identifier for human drugs in the United States.

Rx Norm: A normalized naming system for generic and branded drugs produced by the National Library of Medicine (NLM).

Linking Claims and EHR Data Can Lead to New Insights

Combining EHR and claims data can offer insights that may include a deeper understanding of patient experience, diagnostic evaluations, and physician sentiment about treatment decisions that yield improvements throughout the drug development lifecycle, ultimately leading to improved patient care.

There is a growing acceptance that a claims database linked to data from EHRs can be highly complementary and address some of the known limitations of each data source independently. The depth of the data model that can be created unlocks new use cases and leads to more precise insights.

In essence, by combining claims and EHR data–as is done in many Verana Health Qdata™ modules–life sciences companies gain insights that are greater than the sum of their parts. These insights may include a deeper understanding into patient experience, diagnostic evaluations, multidisciplinary care, and physician sentiment about treatment decisions that yield improvements throughout the drug development lifecycle, ultimately leading to improved patient care.

Outpatient drug prescriptions are one area that illustrate the benefit for combined data. Claims data delivers valuable information regarding if-and-when any prescription was filled by a patient, while EHR data provides information regarding the reason the drug was prescribed, why it was selected, as well as insights on how the patient felt after taking the drug and why they might elect to stop it. The EHR may also have information on certain over the counter medications not routinely captured in a claims database to give a holistic view of a patient's medication profile.

Another illustrative example is a disease such as glaucoma which is managed by ophthalmologists who utilize a combination of prescription eye drops and procedures, as necessary, to manage the disease over time. Typically, data regarding glaucoma patients begins with what physicians record in the EHR at the point-of-care. This EHR data revolves around the patient-physician encounter, and is likely to include symptoms, diagnoses and outcomes, plus information about patient demographics and subgroups. Researchers can then layer claims information regarding prescriptions and fill dates onto the already-strong foundation of EHR data to develop a 360-degree view of the patient. Claims data includes information about which minimally invasive glaucoma surgery (MIGS) procedure a patient has undergone, while EHR data adds further detail about which medical device was used in the procedure from the operative reports. Having a more comprehensive data set can benefit all phases of the drug development lifecycle.

Example: Glaucoma patient journey by data source

Patient Journey	**Screening;** may be determined **glaucoma suspect**	Patient **diagnosed with glaucoma,** assessed for risk, set target pressure	**Treatment #1:** Patient is prescribed topical IOP lowering therapy	**Treatment #2:** Physician performs a SLT laser or MIGs procedure to lower IOP	**Ongoing Monitoring** +/- additional interventions
EHR Data	• Visit Date • Demographics • Ins Coverage • IOP	• Visit Date • Demographics • Ins Coverage • Exam (IOP) • Diagnosis • Risk Factors	• Visit Date • Demographics • Ins Coverage • Medication h/o • Treatment Adherence/ Switching Reason	• Visit Date • Demographics • Ins Coverage • Laser Settings	• Visit Date • Demographics • Ins Coverage • IOP
Ancillary	• Visual Field/OCT	• Visual Field/OCT	• Visual Field/OCT		• Visual Field/OCT
Claims Data	• Ins Payer	• Ins Payer	• Ins Payer • Procedure Code • Prescription Filled	• Ins Payer • Drug Product	• Ins Payer

The size of real-world data sets combining EHR and claims data can provide better insights into health outcomes. Consider, for example, the rapid evolution of therapy options in multiple sclerosis (MS).[3] A number of new MS drugs have entered the market in recent years aimed at two subsets of patients with MS based on disease pattern— one with a progressive course and the other with a remitting and relapsing course. This subcategorization is not currently captured in claims data, but is readily found in EHR documentation. Further, while there are algorithmic ways to use claims data to evaluate if a patient has experienced a relapse of disease such as treatment switching, EHR documentation is more likely to capture the true nature of events at the time of relapse and treatment switching. Therefore, combining the complete therapeutic data from the claims database with the detailed clinical characteristics and outcomes from the EHR database provide invaluable insights into patient diagnoses, treatment patterns, and clinical outcomes in the real world. This type of data can highlight which patients are benefiting from current therapeutic options and for whom treatment is still necessary. Further, high-quality real-world data can be incorporated into health technology assessments, considered for regulatory submissions, and inform guideline development. All of this helps benefit patients by accelerating drug development and advancing routine care.

[4] https://www.amjmed.com/article/S0002-9343(20)30602-1/fulltext

In the Real World: Life Sciences Use Cases for Combined Data

Many life sciences companies have partnered with Verana Health to leverage its quality Qdata™ modules—which often combine EHR and claims data— to inform insights across the drug development lifecycle.

Below are three notable Verana Health use cases for Qdata:

- **Post-approval safety and commercial uptake:** A pharmaceutical company leveraged combined EHR and claims data to monitor post-approval safety and commercial uptake of an anti-VEGF (anti-vascular endothelial growth factor) therapy for patients with age-related macular degeneration. The company performed two studies, one immediately post-launch to study the demographic characteristics of patients who received the new drug soon after approval, and another with six-month safety data after launch, to investigate real-world safety and treatment patterns of patients receiving the drug.

- **Market share tracking for commercial planning and market monitoring:** Another pharmaceutical company used claims and EHR data to build a dashboard tracking the market share of four anti-VEGF drugs across five indications. The dashboard, which is updated quarterly, enables the company to track changes in drug uptake share across clinical indications; forecast future sales; measure salesforce effectiveness by territory; measure medication switches and starts; and identify target physicians and practices.

- **Enable early diagnosis of at-risk patients:** A third pharmaceutical company sought to identify low-risk, early-stage prostate cancer patients to better understand their demographics. With the assistance of NLP technology, this company analyzed claims and EHR data from urologic oncology patients, uncovering important insights into patient demographics, clinical characteristics, treatment patterns and patient outcomes.

Claims and EHR: The Best of Both Worlds

Claims databases are uniquely poised to capture a view of billed care that includes procedures, prescriptions, hospitalization information and other utilized resources, while EHRs provide clarity into diagnosis, severity of disease, patient characteristics and clinical outcomes. When the two data sources are combined, life sciences companies are not only able to broaden the scope of questions they can tackle but build confidence in the evidence they generate.

About Verana Health

Verana Health® is a digital health company pioneering quality in real-world data. The Company operates an exclusive real-world data network of more than 20,000 healthcare providers (HCPs) and 90 million de-identified patients, stemming from its strategic data partnerships with the American Academy of Ophthalmology®, American Academy of Neurology, and American Urological Association. Using its clinician-informed and artificial intelligence-enhanced VeraQ™ population health data engine, Verana Health transforms structured and unstructured healthcare data into curated, disease-specific data modules, Qdata™. Verana Health's Qdata powers analytics solutions and software-as-a-service products for real-world evidence generation, clinical trials enablement, HCP quality reporting, and medical registry data management. Verana Health's quality data and insights help drive progress in medicine to enhance the quality of care and life for patients.

To learn more, visit www.veranahealth.com or contact our team at team@veranahealth.com.