carelon

Technical white paper

Carelon Research's Healthcare Integrated Research Database (HIRD®)

Last updated March 2024

Carelon Research curates the Healthcare Integrated Research Database, or HIRD, as one of the largest, most comprehensive healthcare databases in the US. Experts at Carelon Research apply decades of experience developing and analyzing the HIRD to effectively generate real-world evidence in collaboration with clients.

This technical white paper is tailored for healthcare researchers within life sciences, government, and academia evaluating the HIRD as a resource for their research.

Carelon Research's HIRD®

Bridging diverse health data for comprehensive research insights since 2006

The Healthcare Integrated Research Database (HIRD®) is a proprietary database curated and maintained by Carelon Research. The HIRD contains health-related information for individuals currently and formerly enrolled in commercial, Medicare, and Medicaid health plans located throughout the United States, and it includes data that are eligible for research based upon certain data use permissions.

Carelon Research links enrollment records and healthcare claims in the HIRD to a variety of data sources, including laboratory results data, mortality data, oncology data, electronic health records (EHR), race and ethnicity data, social determinants of health (SDoH) data, and vaccination records to meet the needs of particular studies. The HIRD is updated on a monthly basis and historical data are available dating back to January 2006.

Powering in-depth analyses and real-world evidence across diverse healthcare domains

Researchers at Carelon Research use the HIRD to answer queries in the conduct of methodologically rigorous research.

The HIRD supports real-world evidence generation across a variety of research topics:

- Healthcare costs and utilization
- Cost-effectiveness
- Comparative safety and effectiveness
- Drug utilization and treatment patterns
- Natural history of disease
- Patient and provider insights
- Other research areas

Insurance types in the HIRD and their availability for research

The HIRD includes data from commercial, Medicare, and Medicaid insurance plans. Carelon Research ensures that it has the appropriate data-use permissions and authorization before using 'researchable' insurance plan data contained within the HIRD for client research projects.

Commercial plans

Commercial plans include individuals and their covered family members who are insured as individuals or through their employers. Commercial plans include many large employer groups that have employees located across all 50 states. Employer groups may be either fully insured or self-insured, via an administrative services only (ASO) arrangement.

Carelon Research may only use data from ASO employer groups for research projects if permitted by the ASO group. Pharmacy benefits are available for nearly all fully insured plans and a portion of ASO groups, but some employer groups may choose to obtain pharmacy benefits from a pharmacy benefit management (PBM) company that is not affiliated with the current commercial plan (also known as a pharmacy benefit 'carve-out'). A portion of carve-out pharmacy data may be included in the HIRD if the data are provided to the health plan.

Medicare and Medicaid insurance plans

The HIRD also includes individuals with Medicare Advantage and Medicare Supplemental plans. Most Medicare Advantage plans include a pharmacy drug benefit, whereas a small portion of Medicare Supplemental individuals have a Part D drug plan. Finally, the HIRD includes data from various state Medicaid plans. Although Medicaid plans are included in the HIRD, state-by-state approval is required for use in research studies.
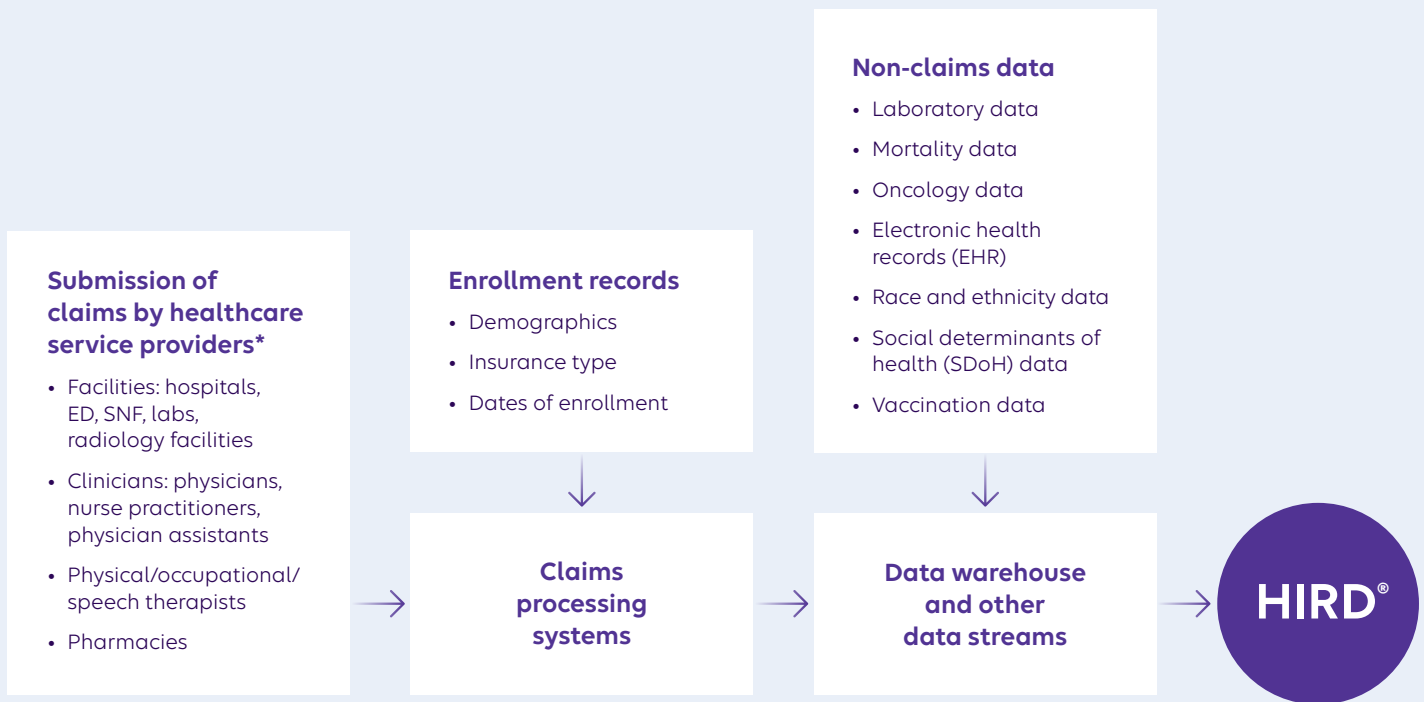
Data sourcing and handling

The data contained in the HIRD come from a variety of sources **(Figure 1)**. Claims submitted by healthcare service providers to health plans and the accompanying individual enrollment records are linked to other external data sources such as laboratory results, mortality data, and oncology data. All data are received by Carelon Research and stored in the HIRD for permissible research use. In addition, Carelon Research uses both deterministic and probabilistic methods to link data in the HIRD to medical records, state cancer registries, the National Death Index (NDI), and other external data sources.

Deterministic linkages are made possible by individual-identifying information contained in the HIRD's enrollment records, such as name, date of birth, sex, and other information, and such linkages are conducted by trained experts. A deterministic linkage to third-party data requires an Institutional Review Board (IRB) Health Insurance Portability and Accountability Act (HIPAA) Waiver of Authorization and appropriate use rights for the data. The HIRD can also be linked using commercially available probabilistic tokenization platforms to expand the reach of the data when deterministic linkages cannot be performed.

Figure 1. An overview of the sources and handling of data contained in the HIRD

Submission of claims by healthcare service providers*

- Facilities: hospitals, ED, SNF, labs, radiology facilities
- Clinicians: physicians, nurse practitioners, physician assistants
- Physical/occupational/ speech therapists
- Pharmacies

Enrollment records

- Demographics
- Insurance type
- Dates of enrollment

Non-claims data

- Laboratory data
- Mortality data
- Oncology data
- Electronic health records (EHR)
- Race and ethnicity data
- Social determinants of health (SDoH) data
- Vaccination data

Claims processing systems

Data warehouse and other data streams

HIRD®

*This list provides examples of entities submitting healthcare claims and is not comprehensive
ED = Emergency department; SNF = Skilled nursing facility

Enrollment records

Enrollment records include a unique member number, a subscriber number (which may be shared among spouses, partners, and/or dependents), information about the type of health plan, and the time period of enrollment in the health plan. An individual can be followed longitudinally across different plans through their unique member number. Enrollment records also include demographic information of the individual as well as geographic (by region) information. Enrollment records are available for all individuals in the HIRD and are updated monthly **(Table 1)**.

Table 1. Detail of information contained in enrollment records, medical claims, and pharmacy claims contained in the HIRD

	Enrollment records	Medical claims	Pharmacy claims
Primary contents	• Member identifier Insurance type • Period of enrollment • Date of birth • Sex • Region • Contact information	• Member identifier • Diagnosis codes • Procedure codes • Admission and discharge dates • Date of service • Place of service • Physician or facility identifier • Plan and member costs	• Member identifier • Prescribing physician • Drug dispensed (NDC) • Date dispensed • Quantity • Days supplied • Drug strength • Plan and member costs
Available since	01 January 2006		
Update frequency	Monthly		
% Adjudicated at 3 months	N/A	>85%	>95%

Maternal health research

The information in the enrollment records allows for linkage of individuals within the same covered unit, typically a family, that in addition to the primary covered individual includes a spouse/partner and eligible dependents. This enables the identification of mother-infant pairs, which can be used to support pregnancy-related research, e.g., studies of treatments received during pregnancy and related risks to maternal and infant outcomes.

Mother-infant linkage is performed by searching for matches on the subscriber identifier and alignment of the mother's recorded delivery date with the infant's date of birth. Approximately 75% of infants can be linked to mothers in the HIRD.[1]

Medical claims

Claims for medical care are submitted by providers for payment of services rendered. Medical claims from healthcare professionals and facilities include:

- Diagnoses
- Services
- Procedures
- Health plan and member costs
- Types of providers
- Site of care

Diagnosis codes

Diagnoses are recorded using the International Classification of Diseases 9th and 10th Clinical Modification codes (ICD-9-CM and ICD-10-CM), with up to twelve diagnosis codes per claim. Procedures are recorded using:

- ICD-9-CM procedure codes
- ICD-10 Procedure Coding System (ICD-10-PCS)
- Healthcare Common Procedure Coding System (HCPCS) codes
- Current Procedural Terminology (CPT) codes

Service information

Medical claims also document service dates, the identity and specialty of service providers, service locations, and service costs. Service locations are classified as occurring in one of four settings:

- Inpatient
- Outpatient
- Skilled nursing facility
- Emergency department

Inpatient and outpatient treatment

Inpatient treatment episodes (days between admission and discharge) are constructed by Carelon Research from multiple claims using a proprietary algorithm informed by service locations and service dates. Telehealth claims for diagnoses and procedures are also recorded.

Outpatient services are reported using the Restructured Berenson-Eggers Type of Service Classification System from the Centers for Medicare and Medicaid Services and updated annually.

Service costs and provider information

Service costs include the amounts charged by the provider, allowed by the health plan, paid by the health plan, paid by the member, and any payments made by another health plan (coordination of benefits). Medical claims also include provider information, such as National Provider Identifier (NPI) and tax identification numbers, practice name, and practice address. This information enables identification of provider specialties, linking providers to a given practice, and determining density of individuals cared for by a provider.

Medical claims updated monthly

The medical claims are updated monthly, and over 85% of claims from all places of service are adjudicated within three months **(Figure 2)**. As a complete medical claims history for a defined time period is important for many research studies, a three-month lag is typically imposed on medical claims data. This practice balances data recency with completeness, while still allowing for access to the most recent data for projects that benefit from this, such as exploring the use or safety of newly approved medications.

Figure 2. Estimated timeliness of medical and pharmacy claim adjudication in commercial claims based upon 2022 data analysis
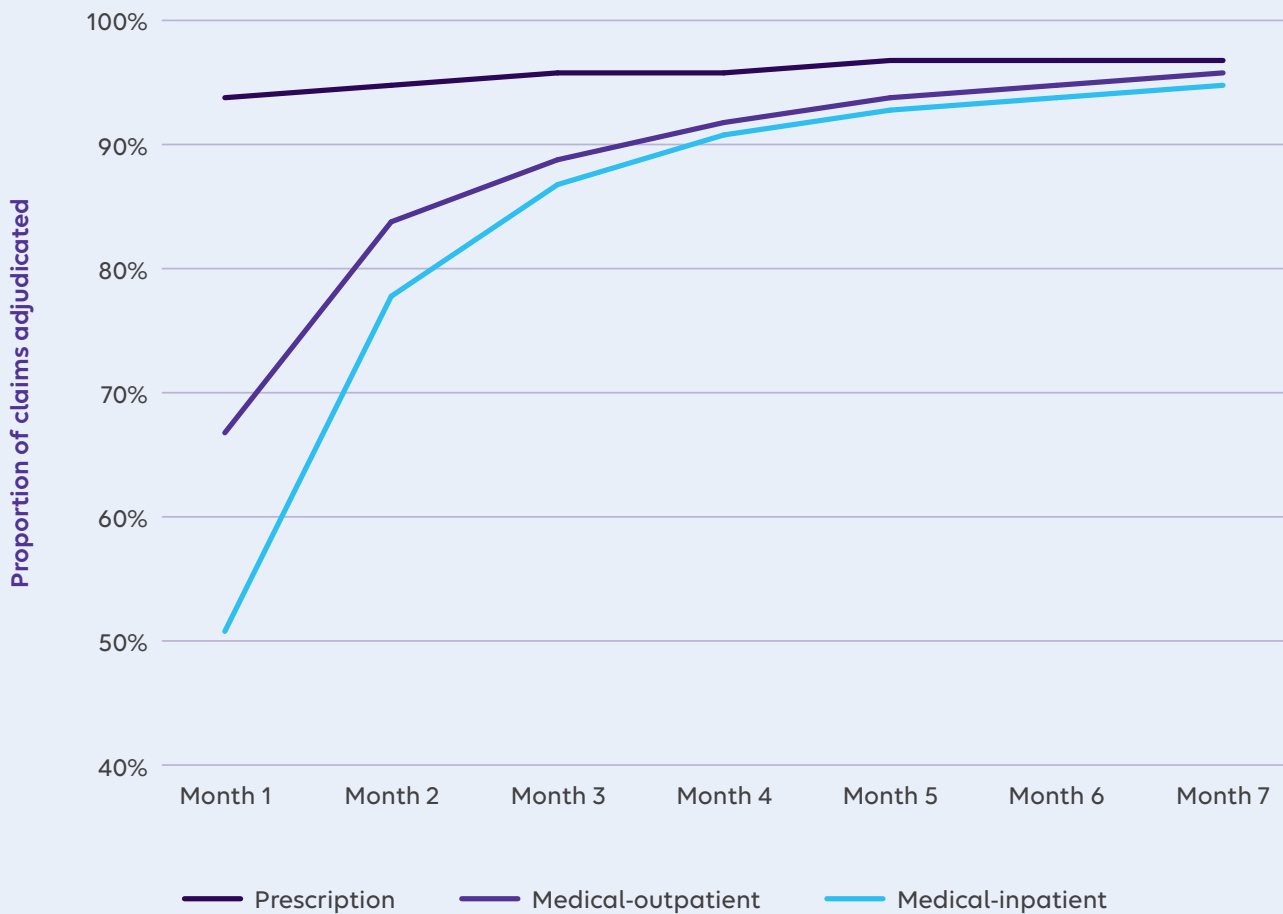
Pharmacy claims

Claims for self-administered prescription treatments are submitted by outpatient pharmacies (including specialty pharmacies). Claims for pharmacy treatments include National Drug Codes for the treatment dispensed, service dates, service costs, prescriber information, location where the prescription was dispensed (pharmacy, mail order service), and information about the pharmacy that dispensed the treatment.

National Drug Codes provide information about the manufacturer, class, strength, package size, route of administration, and dosage form. National Drug Codes are mapped to Generic Product Identifier codes (Wolters Kluwer), which provide further information on the drug group, class, sub-class, base name, name, formulation, and strength.

Costs of prescription treatments are available for all pharmacy claims. Service costs include the amounts paid by the health plan, the individual, and any payments made by another health plan.

Pharmacy claims are available for all individuals in the HIRD with pharmacy benefits. Additionally, pharmacy claims are also available for individuals of some groups who purchase pharmacy benefits separately (carve-out pharmacy). For the carve-out pharmacy claims received, actual costs have been imputed for use in the HIRD. Pharmacy claims in the HIRD are updated monthly. Approximately 90% of pharmacy claims are adjudicated within one month, and approximately 95% are adjudicated within three months **(Figure 2)**.

Components of the HIRD

Laboratory results data

Laboratory test orders and laboratory test results for individuals who receive outpatient laboratory services are integrated within the HIRD. Outpatient laboratory tests ordered by healthcare providers can be identified using procedure codes in individuals' medical claims.

Laboratory test results submitted by certain laboratory providers and electronic health record data sources can be defined using Logical Observation Identifiers Names and Codes (LOINC). A laboratory test's LOINC provides detailed information regarding the test name, specimen source (e.g., serum, urine), how it was measured (e.g., volume, mass), time aspect (e.g., point measurement, 24-hour collection), the scale of the measurement (e.g., quantitative, ordinal, nominal, narrative), and method of collection. Carelon Research has more than 70 of the most commonly used laboratory tests readily available for research in the HIRD, and this can be easily expanded to other labs of interest. Laboratory data are updated monthly.

Mortality data

Mortality data for individuals are obtained from multiple sources, including enrollment records (reason for disenrollment), inpatient claims (discharge status), Death Master File from the Social Security Administration, utilization management data, Center for Medicare and Medicaid Services records, and obituary information processed by third-party vendors. Data from these sources are combined to create a composite mortality variable for research purposes. Carelon

Research has compared the composite mortality variable to the NDI and observed good agreement during the period 2010 to 2018 (e.g., sensitivity and specificity 89%).[2]

While historic mortality data are available from January 2006 to the present, specific sources vary over time. Mortality data in the HIRD are updated monthly.

Oncology data

The HIRD includes oncology data from the health plans' Cancer Care Quality Program (CCQP) integrated with medical claims data. Detailed clinical oncology data for individuals who are undergoing cancer treatment in outpatient settings are captured when a healthcare provider submits a request for preauthorization of a cancer treatment.

Data entered by providers into the CCQP online portal include cancer type, cancer stage, biomarkers and other relevant clinical details (i.e., menopausal status for women with breast cancer), pathology/histology, line of treatment, weight and height, and a metric of functional status (Eastern Cooperative Oncology Group Performance Status Scale). The CCQP data are available in the HIRD beginning July 1, 2014. and are updated monthly. Data were validated in a study published in 2017.[3]

Electronic health records

The HIRD contains both structured and unstructured EHR data for a subset of individuals. The EHR data are obtained from provider network EHR systems, large health systems and clinics, and state-level health

information exchanges. While most EHR data comes from outpatient primary care providers, some also include records from specialists and inpatient providers.

The structured data in the EHR includes anthropometrics (e.g., height, weight), vital signs (e.g., blood pressure), behavioral risk factors (e.g., smoking history), medical history, and medications prescribed. In addition to the coding systems used in medical and pharmacy claims, the structured data in the EHR may also contain Systematized Nomenclature of Medicine (SNOMED) codes (diagnoses, procedures), clinical drugs normalized (RxNorm) codes (treatments), and CVX codes (vaccinations).

Unstructured data in the EHR includes provider office visit notes. These unstructured data can be used to derive structured fields that are readily available for use. Through natural language processing, unstructured clinical data from EHR can be queried to identify clinical information, such as ejection fraction values for patients with a heart failure diagnosis, to create structured fields. These include percent value, percent range, classification for heart failure (reduced, mid-range, preserved), and clinical source.

For individuals for whom EHR data are available, these data may not be available for all periods during which individuals are enrolled in health plans. The EHR data are available in the HIRD beginning January 1, 2010, though the data are most complete after January 1, 2018. The EHR data in the HIRD are updated monthly.

Alternative access to medical records

From the HIRD population identified through claims data, certain research projects may include medical record (both inpatient and outpatient) acquisition and abstraction for capture of clinical information not available in claims data. Through appropriate IRB approvals, Carelon Research can request medical record information of specific patients for clinical data abstraction.

Upon identifying relevant practices and facilities, Carelon Research develops a patient list that includes patient identifiers from enrollment records and facility contact information. Each medical record is then abstracted to obtain the minimum necessary clinical data for each patient for the particular study. A standardized form will be prepared that describes in detail the variables to be abstracted from each medical record. The abstracted clinical data for each patient can then be integrated with their administrative healthcare claims.

Race and ethnicity data

Individual-level race and ethnicity data are obtained from a variety of sources, including self-reported data in enrollment records, EHR data, third-party sources, and algorithms.[4] The combined data sources were validated against internal self-reports indicating good concordance. Race and ethnicity data in the HIRD are classified using Office of Management and Budget standards (i.e., White, Native Hawaiian or Other Pacific Islander, Black and African American, Asian, American Indian or Alaska Native, Hispanic or Latino).

Availability of race and ethnicity data varies by time (recently enrolled individuals are more likely to have race and ethnicity data) and plan type. The race and ethnicity data in the HIRD are updated monthly.

Social determinants of health data

A variety of SDoH data have been integrated into the HIRD.[4] Individual-level information

about urbanicity is derived from the National Center for Health Statistics' urban-rural classification scheme.

The HIRD includes area-level data from the American Community Survey and specifically includes over 50 SDoH variables at the census block group level related to healthcare resource utilization, including educational attainment, income, living conditions (home value, ownership type, crowding), family composition (single parent, number of children per household), transportation, and a socioeconomic status index. The HIRD also includes area-level data from the Food Access Research Atlas with over 140 variables at the census tract level related to food access and availability, urbanicity and rurality, and income. SDoH data are available for up to 95% of individuals starting in 2016.

Vaccination data

To supplement the vaccination data not identified in the medical and pharmacy claims or EHR data, vaccination data from the Immunization Information System (IIS) are also included in the HIRD. Most participating states with IIS immunization data include such data for individuals of all insurance types offered in the state. Vaccination data in the HIRD are updated monthly.

Data use permissions

Carelon Research's access, use, and disclosure of protected health information (PHI) complies with the HIPAA Privacy Rule (45 CFR Part 160 and Subparts A and E of Part 164). To comply with HIPAA, Carelon Research maintains a global Data Sharing Agreement and Business Associate Agreement with the health plans that contribute individuals' protected health information to the HIRD.

Carelon Research does not access, use, or disclose PHI other than as permitted by HIPAA. Limited datasets for research are created when feasible; however, when that is not feasible, Carelon Research may seek to obtain a specific waiver of the HIPAA authorization requirements from an IRB. Carelon Research also takes into consideration other federal and state laws and regulations that might limit use of certain types of data beyond HIPAA limitations, including laws related to substance use disorders and other sensitive medical information.

References

1. Jamal-Allial A, Guzman M, Crowe CL, Gallagher K, Doherty B, Hoffman S, Avula R, Kalloo G, Eshete B. Characteristics of linked and unlinked mother-infant in the HealthCore Integrated Research Database. *Pharmacoepidemiology and Drug Safety*, 2022; 1: 251-2.

2. Sponholtz T, Jamal-Allial A, Vojjala SK, Papazian A, Eshete B, Mahmoudpour SH, Verpillat P, Beachler D. Time trends in performance of different mortality data sources compared to the National Death Index in the HealthCore Integrated Research Database. *Pharmacoepidemiology and Drug Safety*, 2022; 31: 303.

3. Kern DM, Barron JJ, Wu B, Ganetsky A, Willey VJ, Quimbo RA, Fisch MJ, Singer J, Nguyen A, Mamtani R. A validation of clinical data captured from a novel Cancer Care Quality Program directly integrated with administrative claims data. *Pragmatic and Observational Research*, 2017: 149-55.

4. Tonnu-Mihara I, Rojas-Lazic C, Mack M, Eshete B, Stephenson JJ, Grabner M. SA42 Innovative methods for integrating social determinants of health data with administrative claims to facilitate health equity research. *Value Health*, 2023; 26(6): S404–5.

About Carelon Research

Carelon Research (formerly HealthCore), a wholly owned, independently operating subsidiary of Elevance Health, is a healthcare research company committed to improving the future of healthcare with trusted insights powered by evidence. We work with life sciences, government, academia, Elevance Health, and collaborators on a broad range of research solutions focused on generating relevant information needed to help improve health outcomes, lives, and communities.

Learn more about the HIRD
rwe@carelon.com