1. Healthcare Database Information

Humedica (now owned by Optum)

Humedica, which has recently been bought by Optum, is a clinical informatics company that has a database containing information for 18.5 million patients across the US. In 2012 they had 4,956,264 unique patients.

They extract clinical, operational, and financial data from client EHRs and other IT systems across the continuum of care. Data available includes structured data (e.g., diagnoses, procedures, prescriptions, lab results) and unstructured data (e.g., treatment rationale) using natural language processing.

This patient-based EHR database is representative of the United States within 10% of census data. Inpatient, outpatient, and pharmacy claims are available. Data are indexed by patient from the provider EHR records.

Strengths

- Large database of 18.5 million patients, 150 hospitals, and 1,500 ambulatory groups.
- Expected to grow to 87 million patients by 2015
- Longitudinal view of patient health records
- The database is able to link to public data including: procedure/disease codes, US Census data, and mortality reports
- High data stability with provider contracts of approximately 5 years in duration
- Updated monthly with an approximate 45 day lapse between the patient encounter and the data being available in the database
- Natural language processing (NLP) is used
- All diagnostic and procedure codes included: ICD-9, ICD-10, CPT, DRG, HCPCS & J, and EHR

Limitations
Only 60% of inpatient and outpatient claims are linked by patient
Inpatient, outpatient, and pharmacy claims are not currently linked, but will be in early 2013. Details of this, were not available, unfortunately.
Limited long-term care data, information only goes back to 2007 for some patients.
Out-of-network encounters not captured
Cannot collect extra information from patients, physicians, or research sites
Lose 25% of EHR data when integrated with claims
Filled prescription and prescriptions written outside the network are not available
Administrative costs, charges, and payments are not available.
Drug costs, charges, and payments are not available

**OPTUM Insight Clinformatics Data Mart (OPTUM).**

OPTUMInsight is a UnitedHealthcare company. The data is collected from UnitedHealthcare insurance claims and contains data from the entire United States since 2000. In 2011 they had 12,479,069 unique lives.

Demographics are representative in the South and West but less so in the Midwest and Northeast. Data contains inpatient, outpatient, and pharmacy data which may all be linked to each other by patient. Data are indexed by patient using a unique member ID – members that exit and re-enter the insurer’s coverage maintain the same member ID; if they do NOT maintain the same coverage the member ID will change. The database now contains just over 52 million unique patients and is updated bi-annually. 85 percent of claims are processed within 90 days. In addition to the data, In addition to the data, decision support tools, administrative/financial services, and pharmacy solutions.

**Strengths**
- Inpatient, outpatient & pharmacy claims linked by patient
- Size (11+ years of data, >10M lives per year
• Lab results data are available for a portion of the data warehouse ~1/4 of the patients have at least 1 lab test result available
• Secondary data linkages ARE available - Premier (inpatient hospital detail), SSA Death master file, socioeconomic status, and health risk assessment data
• Eleven years are covered by the database (2000-11)
• 85% of claims are posted within 90 days
• Links back to patients CAN be done on a consulting basis. Same for links back to medical records and physicians (can target physicians for participation in clinical trials).
• OPTUM utilizes ICD-9, CPT-4, and DRG codes in its database. ICD-10 codes will be added next year. In addition, CPT, HCPCS, & J codes are collected.

Limitations
• No patient or disease registry data
• No EHR data available
• No text fields readily available
• No disease severity information BUT access to this information is available on a consulting basis. Link to medical records to capture clinical information eg, disease severity/staging. Target patients for participation in surveys and capture patient perspective on disease severity
• CDM lab data is not complete. It includes lab results from several lab vendors, which do not necessarily show a complete representation of lab results for any given member
• While standard CDM (without the Medicare Part D option) does include a sizable population of those over 65 with commercial insurance, this should not be considered typical of the over-65 population as a whole

Additional information regarding the database:
• Optum data include inpatient, outpatient, and pharmacy claims
• Several large populations due to business acquisitions are being added, Medicare Part D (representing 25% of the Medicare population) is available as an option. Medicaid is not available after 2010 and Medicare is not available after 2006 due to contractual reasons. Medicare Advantage data is available through consulting
Unique member ID is generated for an individual. Members that exit and re-enter the insurer's coverage maintain the same member ID.

Facility claims are available for all patients (as applicable) back to 2000. Confinements (stays in a residential facility for persons civilly confined and includes security measures sufficient to protect the community) are built after 2004 for all patients.

Claims data may be linked to patient medical records, patient surveys, and health plans affiliated oncology registry data through ad hoc research projects with OptumInsight.

Data are updated regularly, at a maximum of once per quarter.

An average member month in the data is 27.4 months.

Information Management System (IMS).

Pharmetrics Plus boasts the most diverse representation of geography, employers, payers, providers and therapy areas in the business. They have been active since January 2006 and have 7 years of data available. Pharmetrics Plus is comprised of health plan claims data for approximately 80 million patients with both medical and pharmacy benefits. As of June, they had 26,614,266 unique patients for 2012. For these patients, they have complete enrollment, medical inpatient, outpatient and retail pharmacy claims including cost variables. Patients are nationally representative on a variety of measured, largely commercially insured with a small subset of managed Medicare patients. In addition, ALL 3-digit zip codes in the US are represented. The database is representative of the U.S. Commercially Insured population by age and gender for individuals under 65 years of age.

Strengths

- Demonstrated track record of publishing high-quality health economic and outcomes research studies
- Covers >1 in 3 Americans, 90% of US hospitals and 80% of US doctors
- Availability of 3-digit zip codes
- All US geographies to compare geographic variation
- Pharmetrics Plus has inpatient, outpatient, and pharmacy claims.
- Admitting, secondary and principal diagnosis codes
- Rich benefit info (copays, deductible, on/off formulary, etc)
- IMS has access to data that describes day-to-day management of health care delivery at the task level so that you can document the trail for products ordered directly from manufacturers and then delivered to the healthcare providers within its network.
- IMS can link their data to other IMS or managed assets
- Major Diagnostic Categories are formed by dividing all possible principal diagnoses (from ICD-9-CM) into 25 mutually exclusive diagnosis areas. MDC codes, like DRG codes, are primarily a claims and administrative data element unique to the US medical care reimbursement system. DRG codes also are mapped, or grouped, into MDC codes.

Limitations
- Registry data, as well as EHR data, are not inherent features of Pharmetrics Plus; but may be accessible when linked to another IMS data asset
- Long term care claims may be present for some patients
- The lapse ranges between 45 days and 6 months depending on the type of service and whether one is referencing service or paid date
- Linking back to patients, physicians, and research sites is not an option due to HIPPA privacy rules – other companies (e.g., Optum) have different legal agreements regarding the ownership of identification data and are able to contact patients, physicians, and research sites without violating HIPPA privacy rules since privacy control is maintained within the same company
- Measure of disease severity are not an inherent feature of Pharmetrics Plus; but may be accessible when linked to another IMS data asset
- Cash payments, referral information, and race are not present in the database
- Laboratory and diagnostic test results are not an inherent feature of Pharmetrics Plus; but may be accessible when linked to another IMS data asset

Additional information regarding data content:
• Hospital in-patient records are not an inherent feature of Pharmetrics Plus; but may be accessible when linked to another IMS data asset, although service dates are available for inpatient admissions and ICD-9 procedure codes are present in the data
• Similarly, outpatient data are not an inherent feature of Pharmetrics Plus, but may be accessible when linked to another IMS data asset. Service dates and procedure codes are available for medical encounters
• Death information is not explicitly captured, but may be derived from other data variables in some occasions
• Medical specialty is readily available in Pharmetrics Plus
• Length of stay can be calculated using service date and/or service units
• Location can be derived by POS and Revenue Code
• Service dates are present with inpatient admissions

GE/Quintiles.

This database is made up of de-identified electronic patient records from users of GE Healthcare's Centricity EHR software. With over 33,000 provider contributors, the database consists of over 32 million patient records. The database grows every month with a yearly average growth of ~25-30%. Data consists of all physical findings, lab values, vital signs and all prescriptions. All activities are time and date stamped.

Strengths
• All physical findings, lab values, vital signs and prescriptions are available
• Anonymized EHR database
• Representative of US population based on gender, age, most diseases, and geographic location
• Able to link back to patients and physicians to collect extra information
• Patient history and family history are available
• Text fields are captured in the database
• Size. Over 32 million total lives across 33,000 practices
- ICD-9, CPT, DRG, and some HCPCS & J codes are included in the database.

Limitations
- Registry data, as well as claims data, are not inherent features; but may be accessible when linked to another GE data asset (partnership with Truven Health claims database, Outcome disease registry databases)
- The lapse ranges between 45 days and 6 months depending on the type of service and whether one is referencing service or paid date
- Linking back to patients, physicians, and research sites is not an option due to HIPPA privacy rules
- Cash payments are not present in the database
- Race and ethnicity are missing for 40% of patients
- Managing and programming the data requires experienced staff
- NLP processing of text data is not performed BUT can be performed by an external analyst with access to the data
- Linkage to the National Death Index is NOT currently available, but expected to be possible given a project that required (according to GE/Quintiles staff)

MedMining.

This database is run by the Geisinger Health System and is a collection of data from Geisinger health plans and Geisinger facilities. Geisinger covers 44 counties and 4,187,091 patients in central and northeastern Pennsylvania and was one of the first systems to introduce Electronic Health Records (EHRs). The Health System is composed of over 5 hospital campuses, 75 medical and surgical specialties, 900 physicians, 500 advanced practitioners, 72 primary and specialty clinic sites, 39 community practice sites, over 275,000 members of the health plan (including 50,000 Medicare Advantage members), and over 37,000 contracted physicians/facilities. The database is updated in real-time. Laboratory results and episodic EHR data are available from the past 14 years.
Inpatient, outpatient, and pharmacy data is linked by patient. Data is patient-centric maintaining only one record per patient recording information even if the patient moves in and out of the plan. The biggest advantage to MedMining is their robust use of electronic health records: all EHR data is fully integrated with claims with real-time updates from Geisinger facilities. Commonly utilized data are: EHR data, economic data, claims data, patient and physician survey; and chart reviews. Detailed physician specialty data is available.

**Strengths**
- Inpatient, outpatient, and pharmacy claims
- Size. Contains 14+ years of data and > 4 million patients
- EHR fully integrated with claims
- Linked by patient, not health plan
- Real time updating in the open database
- Medical services cost and utilization data available
- Large Medicare population (18%)
- Extra data licensing options available include: specialty data, claims data, or manual chart review
- Drug costs, payments, and charges are available in the database
- ICD-9, CPT-4, HCPCS and J codes are available in its database.

**Limitations**
- Not representative of US geographically (and probably racially/ethnically too).
- Only contains patients with a Geisinger health plan or received care at a Geisinger facility

**Premier.**

Premier is a full clinical and financial database. Inpatient, outpatient clinics within the hospital system, and pharmacy claims are available from 2000 containing over 96 million patients and 390 million hospital discharges. Inpatient, outpatient, and pharmacy claims are linked by
patient. EHR data is available through chart abstractions if needed and is linked with Optum/Quality Metrics. This linkage can also provide long-term care data.

The database is representative of the US population and is updated monthly with a 90-120 day lag time between the patient encounter and the data being available in the database. The database is a complete census of all US patients in hospitals with no selection bias and no gaps in the data. Each patient has a unique patient identifier which allows for readmission data to be captured. For data that are directly from medical records, those records are linked directly to the data source (ie: medical charts, death certificates, etc). EHR data are fully integrated with claims through chart abstraction. Text items are also recorded in this manner.

**Strengths**

- Representative of United States population
- Inpatient, outpatient, and pharmacy claims.
- EHR connected through regular chart abstractions. This can collect patient reported outcomes.
- Text items recorded.
- Can be linked to Optum/Quality Metrics for claims (drug cost, payment, medical services claims, etc).
- Size. 12+ years and 96+ million patients and 390+ million discharges.
- Updated monthly.
- Captures all patients in the hospital without selection bias.
- Measures of disease severity captured through 3M APR-DRG
- Have a robust data cleansing process including over 1,500 quality assurance checks.
- Contains ICD-9, CTP, DRG, and HCPCS & J codes. ICD-10 codes will be available in the future.

**Limitations**

- 90-120 days elapse between the patient encounter and data being in the database.
- Not a claims database, so medical services payment are not captured, only the actual cost and charges made are captured.
• No drug cost or payment data is captured.
• Limited drug information (does not capture NDC, GPI, or prior authorization).
• Detail on proportion of lives from Medicare, Medicaid, and ESI groups not captured.

**Clinical Practice Research Datalink (CPRD).**

CPRD is a database capturing inpatient and outpatient claims from the United Kingdom’s National Health Service. The online database, also called CPRD GOLD, contains patient registration information and all care events that general practitioners (GPs) have chosen to record in a singular EHR database as part of their usual medical practice. Information held includes records of clinical events (medical diagnoses), referrals to specialists and secondary care settings, prescriptions issued in primary care, records of immunizations/vaccinations, diagnostic testing, lifestyle information (e.g. smoking and alcohol status), and all other types of care administered as part of routine GP practice. Data in the online system are enhanced by the addition of central mortality data (date and causes of death) as well as certain key data from Hospital Episode Statistics (HES- hospitalized patients). Pharmacy claims are also linked to the inpatient and outpatient data. Finally, the following consented data may be available for some patients: clinical trial information, bio-samples, and patient reported outcomes. Data is collected through electronic health records and encompasses over 64 million patients and more than 25 years. The primary care population in CRPD is expected to grow over time. The database contains an open cohort for both practices and patients. All data is linked to patients and physicians so additional information to confirm diagnoses or conduct surveys is possible. Data is indexed based on NHS number, which reduces the likelihood of duplicates in the system.

**Strengths**

• Size. Contains over 25 years of data (since 1987) with 64 million unique patients and 10 million active patients
• Updated monthly
• Follows the patient through National Health Service (NHS) number
- Contains outpatient data for the entire United Kingdom population
- Can link to additional databases, such as NCIN and MINAP
- New line of data for each specialist seen by patient
- Complete medical service utilization data is available including: length of stay, location of care, dates of procedure within an admission, laboratory and diagnostic test results
- Robust cost data available through sub-databases (Drug Tariff, British National Formulary, First Data Bank)
- Lifestyle data available
- Data can be linked back to patients and providers
- Drug dose and strength are available. In the UK, as each new drug enters use, both a generic code and a branded product code are generated for use in EMR's.
- Have two levels of data available based on research needs. Data can also be bought based on a “track” ex: drug track, cost track, niche track, etc

Limitations
- CPRD uses ICD-10 codes, not ICD-9 codes.
- Does not have disease registry information
- Only contains data from the United Kingdom
- Link to inpatient data (Hospital Episode Statistics) is low, only available for 360 practices nationwide
- Does not contain data for patients who are not registered with the NHS, possible that this population might differ from NHS-registered patients

MarketScan.

MarketScan is a research database from Truven Health (formerly Thompson Reuters) with in-depth data of 158 million patients. They have data from the whole United States but are not representative of every region. The database includes inpatient, outpatient, and pharmacy claims that are all linked by patient. Data providers include hospitals, pharmacies, and health plans (including Medicare, Medicaid, and employer-sponsored insurance. Data are indexed by
patient and contains all diseases. This is a closed database with longitudinal care data (information from Medicare Supplement) with over 700,000 patient records in EHR format (Inpatient drug link to Commercial data). Data from hospitals, pharmacies, health plans, employers, and states have been collected since 1995. Patients are identified by a unique ID code that is consistent across services, health plans, and time. Data is linked to the data source.

The MarketScan database contains subject-specific databases that are updated at different times throughout the year. The main MarketScan commercial and supplemental Medicare database is updated quarterly; health risk assessment, health and productivity management, benefit plan, dental, and laboratory databases are updated yearly; multi-state Medicaid database is updated semi-annually; and the hospital drug database is updated monthly. 91% of claims are available five months from the date of service.

Strengths

- Inpatient, outpatient, and pharmacy claims
- Size. 17+ years of data with 158+ million patients. 40 million with employer sponsored insurance, 3.7 million with Medicare Part B and 6.8 million on Medicaid
- Treatment facility and specialist/provider type recorded
- Full medical services costs and payments and full drug costs and payments are available
- Full pharmaceutical information available within one month of dispensing. Prior authorization data may be collected
- ICD-9, CPT, DRG, HCPCS & J codes

Limitations

- Geographic representation is not representative of US distribution.
- Limited EHR data (719,972 patients)
- Limited prescription data
- Data updated between monthly and annually
  - Monthly: Hospital drug database
Quarterly: commercial and supplemental databases
Semi-annually: multi-state Medicaid database
Annually: Health risk assessment, health and productivity management, benefit plan, dental, and laboratory

- Standard updates have a 3 month run-off and annual updates have a 6 month run-off.
  All data is available within 15 months of service with 91% of claims available within 5 months
- Cannot link back to patients and physicians for additional information
- Convenience sample from mostly large employers and hospitals
- Prescription data is less complete in MarketScan than it is in databases capturing employer-sponsored insurance data

The Health Improvement Network (THIN).

THIN is a database of anonymized electronic health records (EHR) of more than 10 million patients from more than 530 primary care practices in the UK. It is not a claims database and is most commonly used in research on epidemiology, drug safety, health economics, outcomes, and drug utilization. Their patient data is representative of the UK based on gender, age, most major diseases, and geographic location.

Unique to this database is the inclusion of anonymized free text comments accompanying patient records. THIN is the database of information collected exclusively from Vision software used in primary care offices. This differentiates it from CPRD, which is a database of National Health Service data. THIN contains mainly outpatient data and inpatient data only for patients already in THIN (eg: their primary care provider uses the Vision software). THIN is a large database with over 400 variables; however, not all fields are mandatory to be entered.

Patients receive a unique patient ID and THIN captures when a patient registers and “de-registers” with the database, but it is unknown if patients who move in and out of the EHR receive a new patient ID or not. The database contains over 10 million records and has the ability to go back and collect extra information from patients, physicians, and research sites.
Strengths

- Anonymized EHR database
- Representative of UK population based on gender, age, most diseases, and geographic location
- Unique patient identifier
- Able to link back to patients and physicians to collect extra information
- Text items recorded
- Size. Over 10 million total lives and 3.7 million active lives across 530 primary care practices
- Drug information included

Limitations

- Only UK data
- Not as representative geographically as the CPRD
- Race/ethnicity data is weak
- Not as established as a quality research database as the CPRD
- No link to secondary care data or between prescriptions and diagnoses (although there is an implied link).
- Limited inpatient claims - contains inpatient data (hospital episode statistics) only for hospital patients whose general practitioner utilizes THIN
- Limited hospital prescribing information
- Limited information on non-NHS care, OTC medications, non-filled prescriptions, hospital prescribing, and lifestyle factors.

RealHealthData

RealHealthData is an outpatient database of patient/doctor transcripts of individual care sessions. The database has been in effect since 2006 and contains nearly 2 million patients from 1,800 practices and is expected to continue growing. Data capture both the health
information of the patient from the patient as well as the observations and knowledge of the provider on a case-by-case basis. The transcripts are coded using natural language processing (NLP) and can be viewed and queried with varying levels of detail. It captures data from the entire United States but it is not documented as to how representative the population is. Inpatient and pharmacy data are only recorded if the physician does so in the transcript. Contains data primarily from general practices but 12 specialty groups, including internal medicine, gastroenterology, nephrology, psychiatry, rehabilitation, and pan specialists, also participate in the database (neurology/immunology were not listed).

Patients can be searched on a variety of variables including: smoking status and frequency, alcohol use and frequency, recreational drug use and frequency, sexual/physical abuse, family history of chronic disease, prior and current medications, medication drug class, prior medical condition, allergies, surgeries (and reason), mental health status/diagnoses.

**Strengths**
- No other database collects data in this way and is a new frontier of data collection
- Data can be easily segmented based on given variables
- Historical data on treatments, care plans, referrals, disease history, surgeries, etc, is easily found

**Limitations**
- Young database (6 years old, 1,735,000 patients, 1,800 practices)
2. Healthcare Database Content

**Humedica.** Humedica’s database contains inpatient, outpatient, disease information (including measures of disease severity), mortality, and claims related information. Among claims information the following is available: type of specialist, treatment facility data, referrals, and physician data.

**Patient Demographics**
Patient age, race/ethnicity, gender, and birth date are all recorded and representative of the US population.

**Medical Services Utilization**
Among medical services utilization the following are available: length of stay, location of care, dates of procedure within an admission, laboratory test results, and diagnostic test results.

**Cost Data**
No medical service or drug cost, charges, or payment data is available.

**OPTUM.** OPTUM’s dataset contains inpatient, outpatient, disease information, and mortality data (either as discharge status or as an optional link to SSA Death master file). OPTUM is also unique in that it contains facility identifier data as well as physician demographics.

**Patient Demographics**

**Medical Services Utilization**
Medical utilization data is fully captures with the exception of procedure dates within an admission. Diagnostic test information is partially available in the claims database, through lab results. Tumor staging information is available on a consulting basis as claims data may be linked to health plan affiliated oncology registry data. Claims data may be linked to patient medical records to capture test results on a consulting basis.

Cost Data
Optum collects “dollar amounts” on their medical and drug costs, including formulary copay data. The main dollar field available is the standardized costs – an estimate of allowed amount. Paid amounts are available on a consulting basis. Other dollar figures included are the charged amount, the coinsurance amount, the deductible amount, the pharmacy copay, and the pharmacy dispensing fee.

IMS. IMS datasets have charge data master (CDM) information including inpatient and outpatient records, procedures, and dates of service. Out-of-network claims are only recorded if a health plan card was submitted. Hospital characteristics are available to describe the treatment facility. The data are longitudinal, with an average duration of member enrollment of two years. Only health plans that submit data for all members are included in the database, ensuring complete data capture and representative samples. Data contributions are also subjected to a series of quality checks to ensure a standardized format and to minimize error rates.

Patient Demographics
Patient demographics, including age and gender, are available. Age can be derived from the patient year of birth. Plus database are generally representative of the national, commercially insured population in terms of age and gender.

Medical Services Utilization
Location and date of care are captured in medical utilization along with 40% of lab tests.


**Cost Data**

All Medical services and cost data (payments and charges) are available with the exception of medical services and drug payments by insurance. The “allowed amount” is an available cost variable and the recommended measure to determine the cost of care. Cash payments are not included in the CDM or medical claims. Patient liability is represented in the copayment amount.

**GE/Quintiles.** Inpatient data is only available through the link with the “TRUVEN HEALTH” data. As this is an EHR database, mainly outpatient data is available including records of visits including dates of service. All diseases are recorded in the EHR.

**Patient Demographics**

Patient age and gender are available. Race and ethnicity may be available in some instances. The format of dates is: mmddyy

**Medical Services Utilization**

Length of stay, location of care, dates of procedure within an admission are available. Laboratory test results are also available. Limited diagnostic test results are available.

**Cost Data**

No cost data available

**MedMining.** MedMining contains inpatient, outpatient, and mortality data. In addition to procedure data, Medmining also has information on measures of disease severity using tools such as pain scales, mini-Mental Status exams, BMD, T-Scores, and FRAX Scores. Unique to MedMining is that have out-of-network claims from care at non-Geisinger facilities. Treatment facility type data is also included.

**Patient Demographics**
The age distribution of patients in their plan is: 0-9: 4%; 10-19: 7.1%; 20-29: 9.7%; 30-39: 9.8%; 40-49: 11.8%; 50-59: 12.4%; 60-69: 9.8%; 70-79: 6.4%; 80+: 17.1%; unknown: 11.86%. The gender distribution is also equitable: 45.4% are women, 44.2% are men, and 10.4% do not have gender recorded.

Medical Services Utilization

The following medical services utilization data are captured: length of stay, procedure “day of service”, laboratory test results, and diagnostic test results. It is undetermined if location of care by day is recorded.

Cost Data

The following medical services costs are captured: medical service “true” costs, medical service charged to insurance, medical service payments by insurance, and medical services charged to the patient. Among drug costs, “true” drug costs are captured. Some measures of drug costs (payments by insurance and payment by patient) are not captured.

Premier. Premier contains inpatient, outpatient, disease, and mortality data. On its own Premier captures all inpatient charge information and the link with Optum captures actual claims. The following charge data master information is available: hospital inpatient records, admission date, discharge date, and procedure codes. The following outpatient data is available: non-hospital patient records, date of service, and procedure codes. All disease are included in the database and severity is measured through the 3M APR-DRG Severity of Illness code. Mortality data is captured.

Patient Demographics

Among patient demographics, age distribution, race/ethnicity distribution, and gender distribution are captured. In general, the format of dates are: month/year.

Medical Services Utilization
The following medical utilization information is captured: length of stay, location of care, dates of procedures within an admission, laboratory test results, and diagnostic test results.

Cost Data
Medical services “true” costs are collected; however, it does not capture claims or reimbursement. Costs are captured through the actual costs and charges according to hospital financials. Drug costs are not available unless Premier is linked to OPTUM.

CPRD. The CPRD data sets contain hospital in-patient records from the charge data master including admission and discharge dates. Hospital procedure codes are not available, although CPRD is building an in-hospital prescribing link in a partnership with IMS. Outpatient patient records are available, although it is undetermined if date of service and non-hospital procedure codes are available. All disease are available as well as mortality data. CPRD has rich information on specialists seen with a new line of data for each specialist the patient has seen. Because CPRD is based on NHS care, patients without NHS care are not included in the dataset.

Patient Demographics
Patient demographics information is collected.

Medical Services Utilization
All medical utilization data (length of stay, location of care, dates of procedure within an admission, and laboratory test results) are collected. It is undetermined if diagnostic test results are available.

Cost Data
Medical service costs are measured in terms of Healthcare Resource Groups (HRGs) which are used in the UK to group similar treatments together to determine overall costs of treatment for population groups. Drug costs are captured utilizing 3 databases: the Drug Tariff, British National Formulary, and First Data Bank.
MarketScan. MarketScan contains the following Charge Data Master information: hospital inpatient records, hospital admission and discharge dates, and hospital procedure codes. Among outpatient services: non-hospital visits, date of service, and non-hospital procedure codes are recorded. All disease and mortality are captured. The following claims-related data is available: specialist/provider type, out-of-network claims (if submitted), treatment facility data, and detailed physician data (physician ID, provider type, service sub-category code).

Patient Demographics
Among commercial databases the age distribution of patients is: 0-17: 26%; 18-34: 24%; 35-44: 17%; 45-54: 18%; 55-64: 15%. The gender distribution is Female: 52%, Male: 48%. The format for dates throughout the database is: mmddyy10

Medical Services Utilization
Among medical services utilization, length of stay and laboratory test results are captured.

Cost Data
In the MarketScan data, the costs charged are not available. However, what is paid by the insurance plan and the patient IS available. In fact, the paid amount can be broken down into the amount paid to the provider by the patient (DEDUCT, COINS, COPAY), the entity that gave MarketScan the claim (NETPAY), and any third party such as Medicare (COB).

THIN. Inpatient information in THIN is connected to hospital episode statistics (HES) on patients in the EHR who get admitted to the hospital. Because THIN is a primary care based EHR, the outpatient data is robust containing non-hospital patient records, date of service, all diseases, and mortality data. Consultation and specialist data is also captured in THIN. Since HES is a separate database that gets linked to THIN, some out-of-network claims are available through this partnership.
Patient Demographics
Demographic data (age, gender) is available, although race/ethnicity data is weak. Information on the patient’s socioeconomic status is also recorded. Data that are linked to HES are considered to have more accurate race/ethnicity data.

Medical Services Utilization
Full medical utilization data (length of stay, location of care, dates of procedure within an admission, laboratory test results, and diagnostic test results) is available.

Cost Data
Only partial cost information is available. Reimbursement and cost information is unavailable.

RealHealthData. Race/ethnicity, gender, and date of birth are collected.
3. Patient Reported Outcomes

**Humedica.** Some patient reported outcomes are available. Symptoms, smoking history and family health history are available. Other information such as patient preference data, satisfaction scores, functional status/activities of daily living, and absenteeism may be available in the notes.

**OPTUM.** Quality of life measures, Patient preference data, Patient satisfaction scores, Functional status/activities of daily living, and Absenteeism (work loss) / productivity data are not readily available BUT is available on a consulting basis.

**IMS.** Quality of life measures, Patient preference data, Patient satisfaction scores, Functional status/activities of daily living, and Absenteeism (work loss) / productivity data are NOT inherent features of Pharmetrics Plus; but may be accessible when linked to another IMS data asset.

**GE/Quintiles.** A large amount of patient reported outcomes are available including: quality of life measures, functional status/activities of daily living, smoking status, symptoms, and assessment of change in disease status. Additional data can be obtained through an additional contract.

**MedMining.** It is undetermined what, if any, patient reported outcomes MedMining reports.

**Premier.** Patient reported outcomes are not available in Premier unless a medical chart abstraction is conducted.

**CPRD.** CPRD contains information on patient reported outcomes but counts for specific types of PRO data were not provided.
**MarketScan.** MarketScan collects information on physical and mental health through self-assessments and data on absenteeism. There are limited quality of life measures.

**THIN.** THIN contains quality of life, lifestyle, and symptom measures. It is undetermined what other patient reported outcomes THIN contains.

**RealHealthData.** Quality of life and lifestyle measures such as tobacco use, depression, and employment are collected.
4. Drug Information

Humedica. Humedica contains drug information for all patients, including: drug name, dose, strength, days’ supply and quantity dispensed (inpatient only), date and time filled, and date written. Both inpatient and outpatient drug information is available. Date dispensed is not available. The database includes NDC, GPI, and prior authorization information. Drug data is available 45 days after the date dispensed.

OPTUM. OPTUM contains drug information for all patients, including: drug name, dose, strength, days’ supply, quantity dispensed, and date dispensed. Both inpatient and outpatient drugs are also available. Prescription data is available in OPTUM within 30 days. GPI (Generic Product Information) code and drug prior authorization/utilization management (step edits, etc.) information is NOT included.

IMS. IMS contains drug information for all patients, including: drug name, dose, and strength, days’ supply, quantity dispensed, date written, and date dispensed. 85 percent of claims are available within 24 hours in the patient-centric database. Additional information regarding data content:

- All patients have a RX benefit at some point in the time period
- Drug name, dose and strength are derived through the NDC code in the reference file provided by Pharmetrics
- Days supply and quantity dispensed are provided by the majority of data contributors
- Date of the pharmacy fill is represented in the service date
- Inpatient drug detail is present for a subset of patients and outpatient drug detail is present for the majority of patients
- Service date represents the date the prescription was filled
- NDC codes are present for retail pharmacy drugs
- GPI information is present in our drug reference files provided to our clients
- Drug prior authorization/utilization management (step edits, etc.) information is NOT included
• Prescription claims are typically the first to be adjudicated by the health plan - depending on the cycle of the data contributor they are often the available within 30 to 60 days

**GE/Quintiles:** Extensive drug detail including brand name and generic name, dosing directions, etc.

**MedMining.** MedMining contains drug information for all patients, including: drug name, dose, and strength, days supply, quantity dispensed, date written, and date dispensed. Both inpatient and outpatient drug information is available. MedMining contains data on prior authorization and utilization management for prescriptions. It is undetermined if MedMining contains NDC or GPI drug codes and if information on prior authorization/utilization management is available.

**Premier.** Premier contains drug information for all patients, including: drug name, dose, strength, days supply, quantity dispensed (most cases), date written, and date dispensed. Both inpatient and outpatient data are available. It generally takes 90-120 days for prescription data to be made available in Premier. Drugs are coded based on the generic name. The pharmacy database does not include NDC or GPI codes. Prior authorization/utilization management information is not captured.

**CPRD.** CPRD contains drug information for all patients, including: drug name, dose, strength, inpatient prescriptions, and outpatient prescriptions. It is undetermined what other drug information is available.

**MarketScan.** Drug information is available for all patients. The following information is collected: drug name, does, strength, day supply, quantity dispensed, and date filled. Both inpatient and outpatient pharmacy information is included. Additionally, both NDC and GPI codes are recorded. Prescription data is available within one month from the date of dispensing. One study has been conducted on prior authorization/utilization management using the MarketScan database.
**THIN.** THIN contains drug information for all patients, including: drug name, dose, strength, days supply, quantity dispensed, date written, and date and time dispensed. A limited amount of hospital prescribing data is available.

**RealHealthData.** N/A
5. Insurance Information

**Humedica.** No insurance information is captured in Humedica. The exception to this are cases in which drug formulary or prior authorization information is entered into the EHR.

**OPTUM.** OPTUM is a dataset directly from UnitedHealthcare and thus claims and insurance information are available. From 2010 forward, the standard database is all commercial. The geographic distribution compares favorably to the US commercially insured population. Medicare Part D data is available as an option, and Medicare Advantage data is available through ad hoc research projects with OptumInsight. Many HMO plan types are represented. The amount the patient pays for medical services/drugs as a co-payment, the amount the patient must pay first before coverage of drug costs (a deductible), and drug formulary information are ALL included.

**IMS.** IMS has limited insurance information available. They provide the amount the patient pays for medical services/drugs as copayment. Additional information related to insurance:

- Limited Managed Medicaid and Medicare patients are available in the data; however, Pharmetrics Plus has a representation of the 65+ population with comprehensive benefits that have not been "labeled" by the data contribute as enrolled in a Medicare Risk (Medicare Advantage) plan. Most patients are from individual health plans and from some large, self-funded employer groups
- Majority of patients are enrolled in a PPO product
- Copayment or coinsurance is available at the claim level if incurred
- Although the amount the patient must pay first before coverage of drug costs (a deductible) is not available, the portion applied to a deductible is available at the claim level
- Formulary indicator is present at the claim level
- Prior authorization status is not included
- Formulary tier information is not available in the data
**GE/Quintiles.** Categories of coverage are available – more specific details are available upon request.

**MedMining.** MedMining is a dataset taken directly from Geisinger health plans, including both commercial and Medicare Advantage plans. Thus, all claims data are available. Approximately 18% of patients have Medicare coverage.

**Premier.** Data on the breakdown of patients on various types of insurance (Medicaid, Employer Sponsored Insurance, etc.) are unavailable. However, all patients are captured regardless of payer type. The following information is not available: patient co-payment for medical services/drugs, patient deductible, drug formulary information, and formulary tier information.

**CPRD.** CPRD is a dataset directly from the UK’s NHS, so all claims and insurance data are available, including drug formulary information.

**MarketScan.** Approximately 40 million patients in MarketScan have employer sponsored insurance, 3.7 million have Medicare Part B, and 6.8 million have Medicaid.

**THIN.** Since THIN is not a claims database, only partial insurance/cost information available.

**RealHealthData.** Formulary information is available.
6. Technical/Other Details

**Humedica**

Data comes in the form of a csv file that is zipped and encrypted before delivery. The file is approximately 250 Gb. Data may be delivered in three ways:

1. Custom study analytics: data delivered in powerpoint slides and excel file
2. Data extract: client receives structured data elements for internal analysis
3. Humedica North Star: custom product delivered via a web-based analytics tool

Data may be licensed in the following ways:

1. Delivered as a single data extract. Client receives all structured data elements for internal analysis or as a Pan Therapeutic file
2. Client receives a custom enterprise license to utilize across the corporation for all therapeutic areas including HEOR, Epidemiology, and Drug Safety

Multiple levels of data quality checks take place. Data quality checks take place at each stage of our build including high level volumetrics of each table, mid-level volumetrics for major items such as level of CPT visit or type of Observation, and a low-level set of 2000+ queries at the disease level to check for consistency across groups and within groups across the month builds. In addition, during our implementation, we conduct field level validation of a sample of patients vs. the source data.

Humedica employs an analytics team that can assist in creating logistic regression models.

**Contact**

Kevin Kelly
1380 Soldiers Field Road
Boston, MA 02135
Kevin.kelly@humedica.com
781-795-6449
OPTUM

Data is 1 TeraByte in size - can be delivered in SAS or pipe delimited text format, delivered in SAS or other de-limited flat file format. Standard DCM or OMOP format available.

Standard license for 12 mo. with 36 mo. historical data, FTE support, online query tool, and one update costs $490,000. Discounted 2 year contract costs $945,000. Discounted 3 year contract costs $1,363,000.

Additional costs:

- Quarterly Updates: $20,000
- Inclusion of Medicare Part D: $60,000
- Ongoing incremental annual fee: $30,000
- Initial fee for data older than 2009: $17,500 - cost for each additional year up to May 2000: $25,000
- Linking Death and SES (up to 3 variables, additional variables $1,000 each) data: $17,500/cohort
- No charge to share the data if UCB provides the data; $5,500 per data cut delivered by OptumInsight

Database consists of adjudicated claims, so much QC occurs during the payment process. From the payment system, the claims go into a company-wide data warehouse with extensive de-identification and quality processes. Optum extracts customer-ready data from the warehouse.

Optum provides analytic consulting services including: oversight, direction, analytics, feasibility, data mining, market research, quality assurance review of each project, quarterly project review. On the research consulting side, they provide full service protocol driven retrospective database analyses, patient surveys, and chart review studies with a focus on HEOR or pharmacoepidemiology (post-marketing drug safety surveillance).

Standard training and support provided:
• Data and analytic specialists on call during business hours.
• Four hour training session per license renewal provided with purchase of the Data Mart.
• Additional hours of training via webex can be arranged.
• Enhanced data training is also available.

Contact
Jennifer Frytak
Director of Business Development
jennifer.frytak@optum.com
952-935-4710

IMS

• Standard data deliverable consists of 1 claims and 2 enrollment files, a proc contents for each file, a procedure, drug and diagnosis reference file and Read Me file specific to the data extract.
• Data are delivered in SAS or flat files.
• A custom URL is provided if hosted access is licensed.
• Licensing options available: clinical extract(s) with a single or multi-use license, single or multi-year licensed.
• Pharmetrics Plus data undergoes rigorous quality controls to ensure unmatched uniformity, quality, and integrity with single data model, 4 level certification, and external review.

IMS HEOR Consultants have deep consulting and pharmacy expertise and robust training in order to ensure insightful and actionable recommendations. They deep experience with a variety of qualitative and quantitative methodologies. Training and support are available via remote or on-site venues.

Contact
Steve Williams  
Director, Global Business Development  
IMS Health Economics & Outcomes Research  
steve.williams@us.imshealth.com  
410-923-0978

**GE/Quintiles.**

*Technical.* Data available in linked text files.

*Research Services.* GE/Quintiles staff can provide guidance on selecting the appropriate dataset, data extraction, analysis, reports, manuscripts, protocols, and ethics/scientific review procedures.

*Cost.* Project-dependent

**Contact**  
Patrick Curry  
Patrick.curry@quintiles.com

**MedMining**

*Technical.* The standard dataset can be complimented with specialty data (Cancer, Rheumatology, pulmonary function test, etc), Claims data (pharmacy, medical, hospital, member), and manual chart review data (staff manual review charts to extract non-discrete data).

*Research Services.* They also provide research staff to assist with studies.

*Cost.* Available upon request

**Contacts**  
Jim Peters  
Chief Executive Officer  
jimpeters@medmining.com
Premier.

Technical. Datasets are typically sent in SAS although other formats can be made available. The data is approximately 700 GB and delivered through FTP site. Data goes through a thorough cleaning process of more than 1,500 quality checks and multiple verification processes.

Research Services. Premier has research and analyst support available. A 3-4 hour WebEx session is the standard training time.

Cost. A minimum one-year license is required.

Contact

Kevin Kozel
kevin_kozel@premierinc.com
570-335-1184
www.premierinc.com

CPRD

Technical. CPRD is a web-based dataset with different levels of data available: Gold Level and Silver Level. Silver Level data is based on cohorts. The Gold level is the full online
database. Data can also be purchased based on track. Examples of tracks include: drug track, cost track, cohort track, niche track, etc.

Research Services. CPRD provides a full-service pharmaco-epi research team for observational, intervention, and clinical studies.

Costs. Costs for data for an individual study. Note that requests for individual studies have to be approved by ISAC before the data can be supplied. A study of 100,000 to 300,000 patients be ~ £60K + the £3-£5K data extraction costs. If HES were included for the same cohort, add an additional £31,500. In addition to costs for HES, there is now a fee of £9,800 for a study involving the cancer registry linkage. There is also a charge for linked MINAP data but this is determined on a study by study basis by MINAP.

### DATA

#### Non-Academic Fee

**GOLD Data set costs**

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 1,000 subjects</td>
<td>£15,000</td>
</tr>
<tr>
<td>1,000 to 50,000 subjects</td>
<td>£35,000</td>
</tr>
<tr>
<td>50,000 to 100,000 subjects</td>
<td>£40,000</td>
</tr>
<tr>
<td>100,000 to 300,000 subjects</td>
<td>£45,000</td>
</tr>
<tr>
<td>Over 300,000 subjects</td>
<td>£60,000</td>
</tr>
</tbody>
</table>

**Additional costs for linked data**

<table>
<thead>
<tr>
<th>Subjects</th>
<th>Fee</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic HES data for less than 100 subjects</td>
<td>£4,500</td>
</tr>
<tr>
<td>Basic HES data for 101 to 10,000 subjects</td>
<td>£7,500</td>
</tr>
<tr>
<td>Basic HES data for 10,001 to 50,000 subjects</td>
<td>£19,000</td>
</tr>
<tr>
<td>Basic HES data for 50,001 to 100,000 subjects</td>
<td>£24,000</td>
</tr>
<tr>
<td>Basic HES data for over 100,000 subjects</td>
<td>£31,500</td>
</tr>
<tr>
<td>Full HES data for less than 100 subjects</td>
<td>£5,000</td>
</tr>
<tr>
<td>Full HES data for 101 to 10,000 subjects</td>
<td>£8,500</td>
</tr>
<tr>
<td>Full HES data for 10,001 to 50,000 subjects</td>
<td>£21,000</td>
</tr>
<tr>
<td>Full HES data for 50,001 to 100,000 subjects</td>
<td>£26,500</td>
</tr>
</tbody>
</table>
Full HES data for over 100,000 subjects £35,000

**Data Extraction Costs** £3,000-£5,000

**Contact**
John Parkinson
john.parkinson@cprd.com

Susan Eaton
susan.eaton@cprd.com
www.cprd.com

**MarketScan.**

*Technical. See above*

*Research Services.* Online training services and manuals are available

*Cost.* Available upon request

**Contact**
Fancesco Colbertaldo
Francesco.colbertaldo@thompsonreuters.com
908-221-0980

**THIN**

*Technical.* Available upon request

*Research Services.* THIN staff can provide guidance on selecting the appropriate dataset, data extraction, analysis, reports, manuscripts, protocols, and ethics/scientific review procedures. THIN also maintains databases in France, Germany, Italy, and Spain.

*Cost.* Available upon request
Contact
Harshvinder Bhullar
Harshvinder.bhullar@cegedim.com
www.cegedimstrategicdata.com

RealHealthData.

Research Services. Data owners can work with clients to query data using refined NLP processing to capture historical outcomes data based on doctor-patient interactions around healthcare utilization and treatments.

Cost. Available upon request

Contact
Manuel Prado
----------------------------
Real Health Data
www.realhealthdata.com
P (415) 251-1066
F (415) 494-2683