**Datasets Frequently Used in Value Research at UCSF**

(for all information and links below, and more, visit healthvalue.ucsf.edu)

**Databases Containing Claims for Payment**

**Medicare claims**: Medicare claims come in a number of forms, including random 5% and 20% samples, a 100% sample (very hard to get), files for specific conditions pre-assembled in the Chronic Conditions Warehouse, and custom data pulls you can request (e.g., all the patients who had cataract surgery in the last 2 years).

*Medicare Claims Link:* <https://www.resdac.org/>

*Medicare Claims Pro Tip:* The 5% and 20% random samples are drawn anew each year, so you cannot use, eg, 10 years of the 20% random sample to actually follow specific individuals over 10 years. However, you can get an “enhanced 5% sample” that will allow this.

**Truven Health (IBM) MarketScan Databases**: These are a set of databases of insurance claims compiled from health insurers, large employers, and government programs. Truven reports that, since 1995, the claims of 240 million unique individuals have been incorporated into the databases. The actual number of patients included varies year to year, but patients can be followed across years and insurers (as long as they stay with an insurer or employer in the dataset). Claims include physician, hospital, and pharmacy claims.

*Truven/IBM link:* <http://truvenhealth.com/markets/life-sciences/products/data-tools/marketscan-databases>

*Truven Health MarketScan Databases Pro Tip:* Providers (e.g., physicians, hospitals) cannot be identified. Furthermore, because different insurers use different identifiers for providers, claims cannot be aggregated up to the provider level (even if using the encrypted identifier) for the providers.

**Optum Insight Databases**: These are a set of databases of insurance claims compiled from health insurers, large employers, and government programs. Optum reports that, since 1993, the claims of 216 million unique individuals have been incorporated into the databases. The actual number of patients included varies year to year, but patients can be followed across years and insurers (as long as they stay with an insurer or employer in the dataset). Claims include physician, hospital, and pharmacy claims.

*Optum Insight link:* <https://www.optum.com/solutions/data-analytics/data/real-world-data-analytics-a-cpl/claims-data.html>

*Optum Insight Pro Tip:* Optum offers seven different “views” of their data, and the choice of view matters a lot. For example, as of March, 2018, you cannot get a view that provides both an indicator for patient death AND a patient geographic location smaller than the state in which the patient resides. Therefore, you cannot, for example, calculate mortality in a ZIP code, city, or county.

**All Payer Claims Databases (APCDs)/Multipayer Claims Databases (MPCDs)**: These are databases compiled by individual state agencies or state-designated organizations that aggregate claims from all or most of the payers (insurers and large employers plus Medicaid and often plus Medicare) in a state. Patients usually can be followed for years, even if they move from employer-based insurance to Medicaid to Medicare. Claims usually include physician, hospital, and pharmacy claims. The APCD Council, the National Association of Health Data Organizations, and the University of New Hampshire maintain a list of which states have APCDs.

*Map Listing States with APCDs:* <https://www.apcdcouncil.org/state/map> .

*APCD Pro Tip:* APCDs capture the vast majority of the citizens of a state in a database that allows them to be followed over time. The exceptions to this are patients without insurance. Because they don’t have insurance, there are no claims filed for these patients and nothing is reported to an APCD.

**Databases Containing Claims Hospital or Emergency Room Discharge Abstracts**

**Vizient (formerly UHC) Database**: Vizient is a company that collects clinical, administrative, and financial data for individual patients from over 200 academic and community hospitals to provide benchmark measures on healthcare resource utilization. Because of the rich data about teaching hospitals, this database is especially used to study care in academic centers. Data is available for adult, pediatric, and neonatal patients. You can easily identify UCSF patients and compare UCSF outcomes with other hospitals across the United States.

*Vizient link:* <https://www.vizientinc.com/>

*Vizient Pro Tip:* Look in the support files to determine which variables are consistently reported across all hospitals. More detailed information is released for UCSF patients compared to other patients from other hospitals in the database. You can also use the online report builder to quickly and easily compare UCSF outcomes with other hospitals of your choice.

**Healthcare Cost and Utilization Project (HCUP) administrative data**: Run by the federal Agency for Healthcare Research and Quality, the HCUP data is a compilation of many state-level databases, such as hospital discharges from Florida or emergency room visits in California. In some states, record linkage files are available that allow you to see if a patient has been readmitted to a hospital or had an emergency room return visit.

*HCUP Link:* <https://www.hcup-us.ahrq.gov/databases.jsp>

*HCUP Pro Tip:* Because there are no outpatient claims, it can be hard to know all of a patients’ diagnoses for risk adjustment or risk stratification.

**Databases Containing Information about Care at the University of California**

**APeX**: APeX is UCSF’s electronic health record, which is a version of Epic. In Epic, the electronic record that nurses, pharmacists, doctors, social workers, and others use in clinical operations is called Chronicles. There is also, however, a separate version of that record, called Clarity, that can be queried for research purposes. To assist with this, UCSF has developed a Research Data Browser that allows you to make requests for data or to download files. It does not provide direct access to Clarity, but that can be requested on a custom basis with IRB approval.

*Apex links:*

·         Access to UCSF Research Data Browser (RDB)

o   <http://myresearch.ucsf.edu/research-data-browser>

o   go down to "Getting Access"

·         Access to UCSF RDB Deidentified flat files:

o   Access is the same as RDB - you get both with one request

o   More information on the flat files:  <http://myresearch.ucsf.edu/research-data-browser-de-identified-export-and-flowsheet-files>

·         Access to UCSF Clarity

o   Requires a service request and (usually) IRB authorization because the data contains PHI

o   Some recharge may be required

o   <https://data.ucsf.edu/data-assets/clinical-research>

·         Access to UCSF Caboodle Data Warehouse

o   Same as "Access to Clarity" - data contains PHI

*APeX Pro Tip:* Note all the items before starting the Research Data Browser request form; these items will make the request form easier to complete.

**UC ReX**: The University of California Research eXchange (UC ReX) Data Explorer enables UC investigators to identify the size of potential research study cohorts across the five UC medical centers. Researchers can conduct interactive searches from patient care activities at Davis, Irvine, Los Angeles, San Diego and San Francisco. Data are derived from both inpatient and ambulatory care settings and have been de-identified. Demographics, diagnoses, procedure codes (via ICD codes), top 150+ lab orders, and a proof of concept of four medications data sets are available.

*UC ReX link:* <https://www.ucrex.org/get-started>

*UC ReX Pro Tip:* Searches return cohort counts rather than data sets. To get the data on the patients found, you then have to contact each institution (as of March, 2018).

**The Most Commonly Used Cancer Database in the US**

**Surveillance Epidemiology and End Results (SEER Cancer Database)**: The SEER Program of the National Cancer Institute collects data on cancer incidence and survival from population-based cancer registries throughout the US. The SEER Program registries routinely collect data on patient demographics, primary tumor site, tumor morphology and stage at diagnosis, first course of treatment, and follow-up for vital status. SEER data can be linked to Medicare claims, the Medicare Health Outcomes Survey, and the National Longitudinal Mortality Study. Custom requests can be made.

*SEER link:* <https://seer.cancer.gov/data/>

*SEER Pro Tip:* Information is really detailed at the time of diagnosis, often including tumor markers and stage, but treatment information is in claims. You can get Medicare claims merged to SEER data, so you can study treatment patterns for cancers affecting older patients. But you cannot get claims for younger people merged to SEER yet.