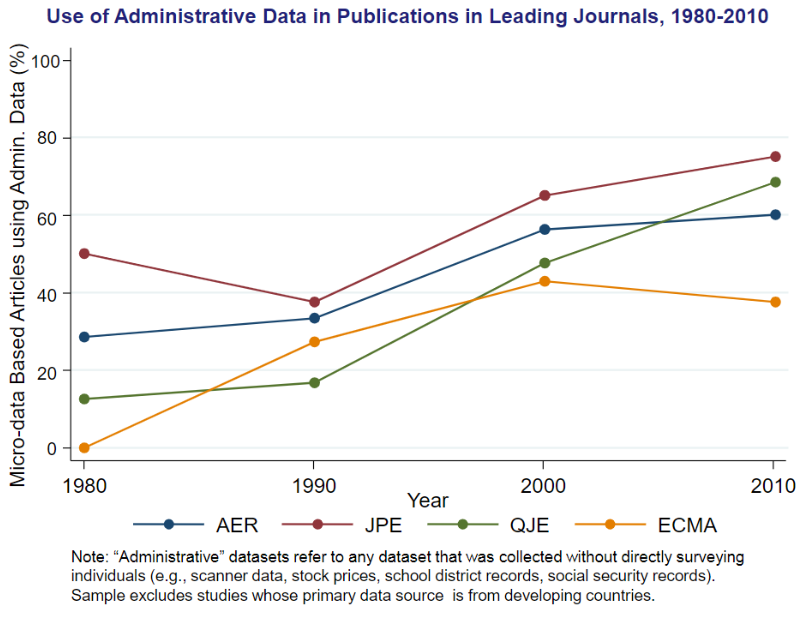
Riley League and Mark Shepard

NBER Health Economics Mentoring Session

May 16, 2024

**Datasets for Health Economics Research**

As you get started with health economics research, it’s important to get good data. Excellent data– often from large administrative data sources, is the lynchpin of applied economics research. As the figure below (from Chetty, 2012) shows use of administrative data in economics has grown rapidly over the last 30-40 years.



Fortunately, health economics has been at the forefront of using big data for applied research on a variety of topics and with a variety of methods. Health economics is also at the forefront of using new data sources – like text data with natural language processing – to study new topics that were previously unmeasurable. This is an exciting time to start in this field!

This document is intended as a brief guide to data sources relevant to health economics research. We hope this can be useful as you start your work. We divide this guide into four areas:

1. **Aggregate health care statistics**: Datasets for getting a big-picture view of the health system and health outcomes
2. **Household survey data:** Microdata on health topics from household surveys
3. **Administrative insurance datasets**: Detailed records on health care use and costs from administrative insurance records
   1. Hospital discharge data
   2. Private health insurance
   3. Medicare and Medicaid data
4. **Other administrative health datasets**

Note that this document focuses on health-specific datasets. Because you will often need non-health data to use as covariates, outcomes, or other purposes, we recommend taking a look at Sebastian Tello-Trillo’s [guide](https://sebastiantellotrillo.com/resources/primer-where-to-find-data) to finding data more generally.

# Aggregate Health Care Statistics

Aggregate statistics on health care relevant issues are widely available in the U.S. and around the world, typically on a public use basis.

In the United States (with which we are most familiar), the key datasets are:

* [**National Health Expenditure Accounts**](https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/nhe-fact-sheet) **(CMS)** – measures of health expenditure, by area of medicine and by state
  + See also the BEA’s [Health Care Satellite Account](https://www.bea.gov/data/special-topics/health-care) data to better measure health spending and value.
* [**Vital Statistics “CDC Wonder” database**](https://www.cdc.gov/nchs/nvss/index.htm)**:** mortality, longevity, births, and other records
  + Note that the disaggregated version of these data can also be used
* **Medicare and Medicaid Public Use data:** <https://data.cms.gov/>
* **Medical Price Statistics:**
  + [Rand Hospital Price Transparency Study](https://www.rand.org/health-care/projects/hospital-pricing.html)
  + [Health Care Cost Institute](https://healthcostinstitute.org/) (HCCI): [Healthprices.org](https://www.healthprices.org/)
  + [Fair Health](https://www.fairhealthconsumer.org/): Medical prices for consumers (mainly outpatient care)
* **Medical utilization statistics:**
  + Hospitals: See [AHRQ’s HCUP project](https://www.ahrq.gov/data/hcup/index.html) for a variety of tools, including [HCUPnet](https://datatools.ahrq.gov/hcupnet/?_gl=1%2A1vr89qe%2A_ga%2AMjAyNjA4NDQ0MC4xNzE1Njk3NTc1%2A_ga_1NPT56LE7J%2AMTcxNTY5NzU3NC4xLjAuMTcxNTY5NzU3NC4wLjAuMA..) and [HCUP Fast Stats](https://datatools.ahrq.gov/hcup-fast-stats/?_gl=1%2A1vr89qe%2A_ga%2AMjAyNjA4NDQ0MC4xNzE1Njk3NTc1%2A_ga_1NPT56LE7J%2AMTcxNTY5NzU3NC4xLjAuMTcxNTY5NzU3NC4wLjAuMA..).
  + [The Dartmouth Atlas](https://www.dartmouthatlas.org/): Medicare use and health outcomes, focusing on regional variations.
* [**HRSA Area Health Resource File**](https://data.hrsa.gov/topics/health-workforce/ahrf): County level data on health workforce, facilities, economics, training, utilization, etc.
* [**State Drug Utilization Data**](https://www.medicaid.gov/medicaid/prescription-drugs/state-drug-utilization-data/index.html) **(SDUD):** state-year level data on drug utilization

For international health data, we recommend starting with the following sources:

* [OECD Health Statistics](https://www.oecd-ilibrary.org/social-issues-migration-health/data/oecd-health-statistics_health-data-en): Good for comparisons across OECD nations
* [WHO World Health Statistics](https://www.who.int/data/gho/publications/world-health-statistics): Good for broader view across nations around the world, including developing countries.

For quick statistics and graphs with international comparisons, [Our World in Data](https://ourworldindata.org/) is an excellent site. For health systems comparisons – often useful for health teaching – the [Commonwealth Fund’s International Health Care Systems Profiles](https://www.commonwealthfund.org/international-health-policy-center/system-profiles) is great (and they have a systems rankings in their “Mirror, Mirror” publication[[1]](#footnote-1)).

# Household Survey Datasets

There are also many large, nationally representative and long-running household surveys that ask respondents questions about their health behaviors, outcomes, and health care consumption, among much else. The following are excellent datasets available on [IPUMS.org](https://www.ipums.org/), which is the easiest way to access microdata. See especially IPUMS Health and IPUMS Global Health

* [National Health Interview Survey (NHIS)](https://nhis.ipums.org/nhis/):
* [Medical Expenditure Panel Survey (MEPS)](https://meps.ipums.org/meps/):
* Census data with health variables:
  + [Current Population Survey (CPS)](https://cps.ipums.org/cps/) for health insurance coverage, self-reported health, and other measures going back decades
  + [American Community Survey (ACS)](https://usa.ipums.org/usa/) for health insurance coverage since 2008.
* [IPUMS Global Health](https://globalhealth.ipums.org/): Demographic and Health Surveys (DHS) from nations around the world, Performance Monitoring and Action (PMA) surveys, Multiple Indicator Cluster Surveys (MCIS).
* [IPUMS Time Use](https://timeuse.ipums.org/) survey may also be relevant for certain questions (e.g. time spent on paperwork).

Other microdata we’ve seen used a lot include:

* [Health and Retirement Study (HRS)](https://www.rand.org/well-being/social-and-behavioral-policy/centers/aging/dataprod.html): Make sure to use the version cleaned by RAND, which is much easier to use.
* [Medicare Current Beneficiary Survey (MCBS)](https://www.cms.gov/data-research/research/medicare-current-beneficiary-survey): Parts of it are publicly available, others require administrative permissions (e.g. to link to Medicare claims)
* [CDC’s Behavioral Risk Factor Surveillance Survey](https://www.cdc.gov/brfss/index.html):
* [CDC’s national Health and Nutrition Examination Survey](https://www.cdc.gov/nchs/nhanes/index.htm) (NHANES)

See also the [NBER’s Public Use Data Archive](https://www.nber.org/research/data?page=1&perPage=50) (and search for “health”).

# Administrative Insurance Datasets

**1. Hospital Discharge Data and Associated Surveys (from HCUP): See** [**https://hcup-us.ahrq.gov/databases.jsp**](https://hcup-us.ahrq.gov/databases.jsp)

* **Patient Samples by care area:** National Inpatient Sample, National Ambulatory Survey Sample, Nationwide Emergency Department Sample (NEDS), National Hospital Ambulatory Medical Care Survey (NHAMCS) (see HCUP website for these)
* **Complete state-level datasets (anonymized patient-level data)** – state-specific harmonized datasets, available for purchase (but check if your institution has them):
  + State Inpatient Database (SID, inpatient admissions)
  + State Emergency Department Database (SEDD)
  + State Ambulatory Surgery and Services Databases (SASD)

**2. Private Insurance Claims**

* **All Payer Claims Databases** (state-specific; see <https://www.ahrq.gov/data/apcd/index.html> and <https://www.apcdcouncil.org/> for summaries).
  + States where we have seen APCDs used for research: Massachusetts, New Hampshire, Oregon, Colorado, Utah, Vermont, Maine.
  + Map of states that have APCDs (at least on paper): <https://www.apcdcouncil.org/state/map>
* **Claims Aggregators:** 
  + **IBM/Truven Marketscan:** claims for a large set of self-insured employer-sponsored plans for employers. (This is expensive, but the NBER and many universities have a license.)
  + [**Health Care Cost Institute**](https://healthcostinstitute.org/) **(HCCI):** Health claims from a consortium of large insurers, available for multiple business lines, but employer-sponsored plans are most commonly used. Expensive, but some universities have a license.
  + [**United Optum Database**](https://www.optum.com/en/business/life-sciences/real-world-data/claims-data.html)**:** Health claims from UnitedHealthcare and Optum insurers across several business lines. Also expensive.
  + [**Inovalon**](https://www.inovalon.com/products/data-cloud/real-world-data/)**:** Claims from providers that use Inovalon IT systems

**3. Medicare and Medicaid Administrative Data**

* [**Traditional Medicare Claims**](https://resdac.org/)
  + [Public Use Files (PUF)](https://data.cms.gov/): non-identifiable, but do contain provider-level data, including information on payments received from Medicare and detailed facility information
  + Limited Data Set (LDS): beneficiary-level information, but no direct identifiers; between PUF and RIF in detail and use restrictions
  + Research Identifiable File (RIF): claim-level data for Medicare Parts A and B for all (or a random sample of) Traditional Medicare beneficiaries
* **Medicaid Claims**
  + Often best to work with a single state because of differences in program administration. (See e.g., [Geruso, Layton, Wallace (AEJ Applied 2023)](https://www.aeaweb.org/articles?id=10.1257/app.20210843) using NY data)
  + [MAX Data](https://www.cms.gov/data-research/computer-data-systems/medicaid-data-sources-general-information/medicaid-analytic-extract-max-general-information): Medicaid claims for a large collection of states that report data

Using administrative datasets often involves a more onerous process for accessing and analyzing the data than aggregate or survey data. Getting access to identifiable data often requires the following steps:

1. **Institutional Review Board (IRB) approval** – If you are accessing identifiable data, your institution’s IRB will need to determine whether you are engaged in human subjects research and if so, approve your plan for that research to minimize the risks to those you are studying.
2. **Data use agreements (DUA)** – The organization providing you with data usually puts restrictions on how the data can be used. Sometimes these restrictions will not be binding and forbid things like attempting to identify de-identified data or using the data for commercial purposes. Other times, the restrictions will require you to specify the research questions you will be asking and how you plan to answer them, restricting you to using the data only for those purposes. DUAs will also often spell out how the data can be stored and accessed and how any output will be reviewed ensure the privacy of the people you are studying. All of these requirements will need to signed off on by your home institution as the DUA is a legally binding contract.
3. **Paying for access** – Administrative datasets are often expensive. For example, the cost to access the HCCI data as a student is $15,000 per year, and the cost to access the VRDC housing the Medicare data is $18,000 per year. However, your institution may already have a license to use the datasets that allows new projects to be added at substantially lower cost. The NBER has a few of these datasets, so if you have an advisor who is an affiliate, they can often get you access.
4. **Waiting** – The process above can often take quite a while to play out. For example, it often takes up to a year from applying for access to the Traditional Medicare data to being granted access.

# Other Administrative Health Datasets

The US health care system is complicated and requires many different entities to keep track of the health care that is rendered to patients and the payments made from enrollees to insurers and insurers to providers. Unlike in other industries, these records are plentiful and can often be used for research. A good rule of thumb for finding administrative data is to ask what entity would need to keep detailed records of the interactions you’re interested studying in order to accomplish its organizational mission. A sampling of other administrative health datasets that we’ve seen used for research include the following:

* Healthcare Cost Report Information (HCRIS)
* American Hospital Association (AHA) Annual Survey
* Veteran’s Health Administration (VHA) Data
* Military Health System (TRICARE)
* Electronic Health Records
* Remittance Data
* [ARCOS](https://www.deadiversion.usdoj.gov/arcos/retail_drug_summary/arcos-drug-summary-reports.html) Controlled Substances Transations

Even more administrative data resources can be found in newsletters from the American Society of Health Economists (ASHEcon) [here](https://www.ashecon.org/newsletter/newsletter-issue-20194/hospital-financial-characteristics-datasets/), [here](https://www.ashecon.org/newsletter/newsletter-issue-20201/data-resources-on-medical-providers/), and [here](https://www.ashecon.org/newsletter/newsletter-20211/data-resources-on-health-care-encounter-data/).

All of the administrative datasets discussed so far rely on data from the US, but there is excellent administrative data available on health care systems around the world. Because many health care systems are more heavily financed by public payers, administrative data can often link health care claims to other administrative data, including birth and death records, tax records, educational data, and much, much more. Accessing these data from the US can sometimes be difficult and often requires working with a collaborator in the country in question.

1. Note, however, that these rankings are biased against market-based systems (in the Bismarck model used in Germany, Switzerland, etc.) and towards public systems with low cost sharing (in the Beveridge model used by the UK, Canada, and others). There have been other rankings, including this one in the [NY Times by health economists](https://www.nytimes.com/interactive/2017/09/18/upshot/best-health-care-system-country-bracket.html). [↑](#footnote-ref-1)