Data Hub

Table of contents

Welcome

The Center for Health Data Sciences **Data Hub** is a university-wide consultation service designed to help faculty and staff navigate the fast-moving world of **data science, machine learning, and AI**. Whether you're just getting started or exploring advanced analytic strategies, the Data Hub is your **first stop for expert advice**.

What We Offer

Our team of data specialists provides short, targeted consultations to help researchers:

- Discover and evaluate relevant **data sources**
- Understand **data access pathways** and ethical considerations
- Strengthen **analytic strategy** and research design
- Facilitate **interdepartmental collaboration**
- Triage complex needs to the right people across BU

These 30-minute sessions are offered at no cost and are open to faculty and staff across BU.

How It Works

1. **Submit a Request**: Use our consultation request form or email us at chds@bu.edu
2. **We'll Schedule a Session**: A data expert will meet with you virtually or in person
3. **Get Connected**: If your request needs more follow-up, we'll refer you to the right partner in CHDS, BEDAC, or beyond

We use a centralized request system to track usage trends, reduce email overload, and continuously improve service.

Topics We Cover

Our team can help with questions related to:

- Sourcing internal and external datasets
- Data licensing and permissions
- IRB and data use agreements (DUAs)
- Matching your research question to the right method or tool
- Building collaborations across BU

Learn more: Data source inventory

Strategic Benefits

The Data Hub isn't just about problem-solving—it's about enabling research at BU to reach its full potential. The Hub helps:

- **Break down silos** across schools and research centers
- **Strengthen proposals** by improving data and methodology plans
- **Save time**, letting investigators focus on analysis instead of logistics

The BU Data Hub is powered by CHDS and BEDAC at the BU School of Public Health, with support from research and IT partners across the university.

Part I

Data Source Inventory

> **Warning**
>
> filled in by CHAT-GPT, need to confirm but honestly looks pretty good

This table provides an overview of population and health-related data sources commonly used in epidemiology, health services, and social science research. It includes a brief description, key attributes, access requirements, and relevant links or notes.

Data Source	Description	Population/Coverage	Access & Restrictions	Notes/Links
MarketScan	Claims data from commercial insurers, Medicare & Medicaid enrollees	National; insured populations (2010–2022)	Licensed via DUA with Truven	Truven/IBM MarketScan Overview
Optum	De-identified claims and EHR data from Optum clients	~100M+ covered lives; national	License required	Often used in comparative effectiveness and pharmacoepi research
CMS Medicare	Medicare Parts A, B, D claims and enrollment files	US adults 65 and disabled populations	ResDAC request and DUA	CMS Data Navigator
NHANES	National Health and Nutrition Examination Survey	National, representative sample	Public + restricted access tiers	NHANES
ACS (American Community Survey)	Ongoing demographic, housing, social, and economic data	National, sub-county level (tract/block)	Public	ACS Data
NSDUH	National Survey on Drug Use and Health	US residents aged 12	Public microdata via SAMHDA	NSDUH
HRS	Health and Retirement Study	US adults 50, longitudinal panel	Public + restricted files	HRS

Data Source	Description	Population/Coverage	Access & Restrictions	Notes/Links
All of Us	National cohort with EHRs, biospecimens, and surveys	Diverse US cohort, oversampling minorities	Controlled tier via Researcher Hub	All of Us Data Browser
BRFSS	Behavioral Risk Factor Surveillance System	National; state-level behavioral data	Public	BRFSS
MEPS	Medical Expenditure Panel Survey	US households and medical providers	Public	MEPS
NSAF	National Survey of America's Families (historical)	Cross-sectional (1997–2002)	Public (retired, archived)	Used in Medicaid and CHIP policy research
Vital Statistics (NCHS)	Birth, death, fetal death, marriage, and divorce data	National and state-level	Public (some restricted use)	CDC Vital Stats
SEER-Medicare	Linked cancer registry and Medicare claims data	Cancer patients 65	Application through NCI/ResDAC	SEER-Medicare
HCUP	Healthcare Cost and Utilization Project	Inpatient, ED, ambulatory visits	State-specific, requires purchase	HCUP

MarketScan

MarketScan is a proprietary claims dataset that provides longitudinal patient-level data on healthcare utilization, expenditures, and outcomes. It includes data from commercial insurance plans, Medicare, and Medicaid populations across the United States. The dataset is widely used in health economics, pharmacoepidemiology, and outcomes research.

Data Dictionary, Metadata, Years Available

- **Years Available**:

- **Geography**:
- **Population**:
- **Documentation**:
  - Download Data Dictionary (PDF)
  - Download User Guide (PDF)

Data Use Agreement (DUA) and Abstract

Access to MarketScan data is governed by a Data Use Agreement (DUA) between Truven Health Analytics and Boston University. Please contact the Data Steward (email below) for more information.

Sample abstract for DUA request: > This project aims to examine the impact of medication adherence on hospitalization rates among patients with chronic conditions using MarketScan Commercial Claims and Encounters data from 2015–2021.

Submit a Project Proposal

You can submit your proposal using the REDCap form below:

Current Projects Using MarketScan

- **Project 1**
  Author

Past Projects

- **Project 1**
  Author
  Final Report (PDF)

Related Publications

-

Code Snippets to Get Started

```
# Load claims data
claims <- fread("marketscan_claims_2021.csv")

# Summary by diagnosis
claims[, .N, by = .(diagnosis_code)][order(-N)]

# Join with enrollment data
merged <- merge(claims, enrollment, by = "patient_id")
```

Staff

If we want to talk about CHDS staff, could do it here