Leveraging Health Insurance Claims Data to Complement Data from Federal Surveys

In the wake of a dynamic environment for federal data infrastructure, state health policy professionals have grown concerned about whether they will continue to have access to data they need for evidence-based policymaking. Uncertainties include the potential cancellation or scaling back of critical surveys, reduced federal survey staffing that could undermine data quality or technical assistance, and the possibility that public data releases may be curtailed or restricted.

To identify strategies to fill potential data gaps caused by federal survey disinvestments, the State Health Access Data Assistance Center (SHADAC), as part of its State Solutions for Health Data Continuity project funded by the Robert Wood Johnson Foundation, conducted discussions with 46 diverse experts and stakeholders, including state data analysts, survey data collectors, researchers, and foundations. During these discussions, SHADAC was able to gather insights on people's top concerns about emerging and potential gaps in federal surveys, and explore various potential alternatives and strategies for filling state-level data gaps.

We have created a series of "Strategy Briefs" that review a selection of strategies or "alternatives" for filling federal data gaps, identified both through these discussions and from SHADAC's expertise.

This brief provides a high-level overview of the strategy: **Leveraging Health Insurance Claims Data to Complement Federal Surveys**. Along with a general description of this gap filling strategy, SHADAC experts will also summarize identified advantages, disadvantages, and implementation considerations for data users and other interested parties.

Complement, not Substitute

A recurring theme across our stakeholder discussions was the **potential for non-survey data to fill emerging gaps**. Important to note: This strategy proposes **using non-survey data as a *complement to*, rather than a *substitute for*, federal survey data**. While both individuals we spoke to as part of this project and SHADAC experts agree that non-survey data cannot replace federal survey data's standardized, population-wide view that enables national benchmarking and cross-state comparability, it's also universally agreed that non-survey health data sources could **help fill gaps** and **complement federal health data** that is collected and released.

Sometimes referred to as "real-world" data due to their basis on actual patient encounters with the healthcare delivery system, access to non-survey health datasets has greatly improved over the last two decades due to advances in health information technology, increased consumer demand for health information, legislation to promote and protect electronic exchange of health information, and digitizing health program data.

While many types of health-related non-survey data exist (i.e., hospital discharge data, disease registries, etc.), in our work and through the informative expert conversations, SHADAC has identified two non-survey data types that could be particularly effective in helping to fill emerging health data gaps: **Health Insurance Claims Data** and **Electronic Health Record Data**. This brief focuses specifically on **leveraging health insurance claims data as a gap filling strategy**; a second brief focuses on the strategy of using Electronic Health Record (EHR) data.

Strategy Description

Health insurance claims datasets are service-level data based on information submitted by providers to payers for reimbursement. This data includes information on healthcare utilization, access, and cost across a wide variety of medical care settings (i.e., outpatient, inpatient, emergency), and often also includes data from various service categories such as pharmacy, imaging and radiology, laboratory, and more. While not created with the intent for public health or health policy monitoring, evaluation, or research, claims data can offer both a comprehensive and granular view of healthcare utilization at the state-level. There are also a number of different types of databases that compile large amounts of health insurance claims data, such as private databases (usually compiling claims from national, commercial insurers), state All-Payer Claims Databases (APCDs) (compiling health insurance claims from a wide array of payers operating in a particular state), and data and statistical resources based on Medicare and Medicaid claims produced by the Centers for Mediare and Medicaid Services (CMS). These claims databases create powerful and large sources of information that could be useful in helping to fill federal health data gaps.

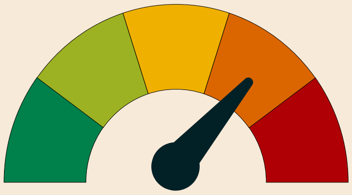
Leveraging Claims Datasets: Strategy Advantages

- **Comprehensive Service Detail**
  Claims datasets provide granular information on billable medical services across diverse care settings (inpatient, outpatient, emergency). They often integrate pharmacy, durable medical equipment, and dental claims, offering a holistic view of a patient's utilization.
- **Large Sample Size**
  Claims datasets typically include all encounters for a specific insured population. State APCDs are particularly powerful because they include claims from multiple providers and payers over time.
- **Low Patient Burden**
  Datasets are based on billed services delivered to patients, not patient self-reports.

Leveraging Claims Datasets: Strategy Disadvantages

- **Population Gaps**
  Claims datasets exclude the uninsured by default. Additionally, state All-Payer Claims Databases (APCDs) often lack data from self-insured employers (due to ERISA exemptions) and federal programs like the Veterans Administration or Indian Health Service.[1] Private databases are typically based on data from large, national commercial insurers. Because these datasets are proprietary, the exact population covered is often opaque.
- **Limited Socio-Demographic Detail**
  While claims data typically include basic demographics such as age, sex, and geography, data on race/ethnicity, language, sexual orientation, and gender identity is limited. In addition, claims databases rarely capture comprehensive social determinants of health, such as income, education, employment, housing stability, or food security.
- **High Barriers to Entry**
  Access is often restricted by high licensing costs or stringent state-level privacy regulations. Additionally, the sheer volume of data requires sophisticated storage infrastructure and advanced technical expertise.

Resource Intensity



Medium to High

Actors

State data collectors

Policymakers

Vendors (data aggregators)

Researchers

Timeline

Short-term stopgap for population health or cross-state analyses

Possible long-term alternative for within-state analyses

Claims Data to Study Healthcare Utilization, Access, and Cost: Examples and Implementation Considerations

Healthcare claims data can be an effective alternative to better understand healthcare access, utilization, and spending at the state level for insured populations. While the federal Medical Expenditure Panel Survey – Household Component (MEPS-HC) collects data on all individuals' utilization and spending on healthcare services, regardless of payer status, it does not support state-level estimates for all 50 states.

Almost half of the states in the U.S. have stood up All Payer Claims Database (APCDs) to house and use data on paid healthcare claims from most public and private health insurers in their state. Common use cases of this data include public reporting of healthcare utilization and prices, assessments of annual benchmarks to limit growth in healthcare spending, and studies to monitor spending on patient care considered "low-value".[1] States without APCDs have relied on data from private organizations, such as the Health Care Cost Institute (HCCI), OptumInsight, Truven Health, and others, which aggregate multi-payer claims data from across the U.S. in database products for purchase and use. For example, the State of New Jersey's Office of Health Care Affordability and Transparency relied on datasets from the HCCI to study healthcare spending trends for residents with employer-sponsored insurance as they worked to promote cost transparency and inform ways the state might reduce spending growth.[2]

The Centers for Medicare and Medicaid Services' (CMS) Transformed Medicaid Statistical Information System (T-MSIS) is another claims dataset that could serve as a stopgap to federal health surveys like the National Health Interview Survey (NHIS) and the Behavioral Risk Factor Surveillance System (BRFSS) to study the Medicaid population and conduct cross-state analyses. T-MSIS harmonizes state-level data on inpatient, outpatient, long-term care, and pharmacy claims into a comprehensive database that facilitates national analyses as well as cross-state and subgroup comparisons.[3]

CMS' investment in the Data Quality Atlas (or DQ Atlas) provides data users with detailed information on the accuracy and reliability of Medicaid and CHIP variables available in the T-MSIS analytic files. As a federally produced data source, there is some concern about whether the federal government will continue to invest in these data and associated resources like the DQ Atlas. However, T-MSIS seems like a promising alternative and complement to survey data for monitoring the Medicaid program, particularly as it undergoes major changes related to H.R.1. For example, T-MSIS could be used to produce 50-state summary data on changes in healthcare utilization post-H.R. 1 implementation.

While a rich data asset, leveraging claims data for state health policy purposes requires significant time, resources, and skill to establish data use agreements, refine scope, assess the quality of variables of interest, and analyze and report findings. Technical assistance can be a valuable resource to help state data users navigate the myriad of variables available in these datasets and the complexities associated with intra-state and cross-state analyses for monitoring healthcare access, utilization, and cost.

States often lack the in-house capacity to access and use these data in-house and rely on vendors, such as research organizations or data aggregators, to clarify use cases, identify inconsistencies across payer submissions, and conduct analyses. Notably, some vendors are now working to address the lack of standardization across different state APCDs. For example, Onpoint is developing dashboards for behavioral health using APCD data from five states to enable comparative analysis and to inform policy action. Claims data could also be linked to survey data to reduce respondent burden.

Continued monitoring, experimentation, and technical support for data users are essential parts of implementing data gap filling alternatives and strategies; SHADAC plans to continue research and provide technical assistance in this area—contact us with questions or feedback.

Interested in learning about other alternatives and strategies for filling potential data gaps? Find and read the other iterations of SHADAC's Strategy Briefs here.

---

[1] Blewett, L., Mac Arthur, N., & Campbell, J. (2023). The Future of State All-Payer Claims Databases. *Journal of Health Politics, Policy and Law, 48*(1), 93-115. https://doi.org/10.1215/03616878-10171104

[2] Office of Health Care Affordability and Transparency. (2026, January). *Health Care Spending Trends for New Jersey Residents with Commercial Insurance, 2017–2022*. New Jersey Department of Health. Mathematica. https://www.nj.gov/health/about/documents/ohcat/NJHARTProgramCommercialCostDriversAnalysisReport_2017-2022.pdf

[3] Nelson, Q., McGlave, C. (2024, October). *Medicaid Claims Data: Using T-MSIS for State Health Outcomes Research*. SHADAC. https://www.shadac.org/publications/medicaid-claims-data-t-msis-research-brief