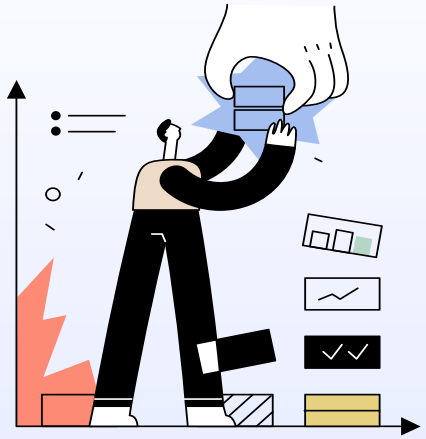
Secondary Data Sources for Nurse Practitioner (NP) Researchers and Students

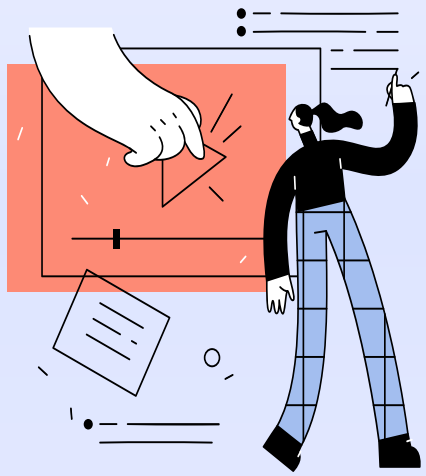
American Association of NURSE PRACTITIONERS®

This guide provides suggestions for finding and assessing **secondary data sources** (Section I) and a **curated collection of datasets** relevant to NPs and NP research, along with links and example studies (Section II).



Secondary data refers to data that were originally collected for a purpose other than addressing a current research question — such as administrative claims, electronic health records, national surveys or workforce datasets. These data are often collected by government agencies, health systems, insurers or research organizations and made available for analysis.

Secondary datasets may seem complex or difficult to access, but NP and student researchers should consider exploring them before turning to primary data collection methods like original surveys. Compared to primary data sources, secondary datasets often include larger sample sizes, broader geographic coverage and more diverse populations. They can also help researchers avoid common challenges with survey-based research, such as low response rates, limited generalizability and high data collection costs.



Although secondary data come with their own limitations — such as restricted variables, data use agreements or limited control over how data were collected — they are a powerful and efficient resource for studying clinical practice, workforce trends, health outcomes and policy-relevant questions in NP research.

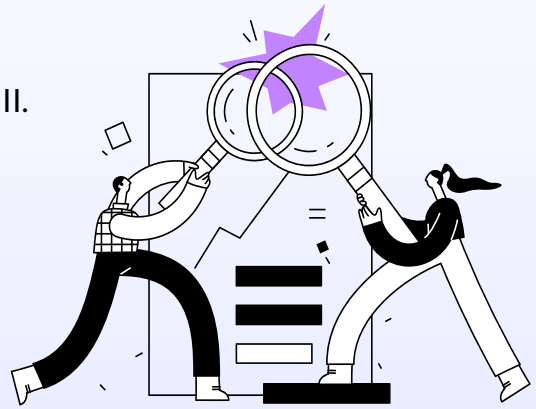
I. Finding and Assessing Secondary Datasets

Step 1: Start with the research question

Develop a testable, focused research question before looking through datasets. Your research question should determine the characteristics of the secondary data you want to analyze and how you analyze the data. From the question, develop your analytical plan. Consider the measures needed to answer your question as well as important control variables.

Step 2: Identify possible datasets for a closer look

Start with datasets used in published literature on your topic or read through the options in the tables in Section II.



Step 3: Review codebooks and data documentation

Codebooks and other materials detail the data collection methods and the items (i.e., survey questions) used to collect data. Pay particular attention to:

- Time structure of the data (i.e., use longitudinal data for a longitudinal question).
- Where the data came from (i.e., a survey, claims or other administrative sources).
- Sample size and characteristics, including the need for sample weights to achieve generalizability.
- Updates, revisions or instrument changes that would make it difficult to compare across time periods.

It is easier to review a codebook than a dataset. The data may contain hundreds of variables and thousands of observations that may be overwhelming. Ensure that the variables as described in the codebook align with your research question and capture the concepts you want to study and control variables you need. After selecting relevant measures, check that data collection methods are reasonable and valid. For example, check the wording of the survey questions yielding the measures you plan to use.

Step 4: Secure permissions and approvals

It is a best practice to seek Institutional Review Board (IRB) approval or exemption for all research, even when a project will likely qualify as exempt from the need for human subjects protections. If your institution does not have its own IRB, you can search for independent IRBs here:

Association for the Accreditation of Human Research Protection Programs.

Be aware of any required data use agreements to use secondary data. Some datasets require application forms, fees or both. Ensure you can safely store data on secure servers and comply with anytime limits on access or use. Review publishing requirements, including data citations and restrictions on dissemination.

Step 5: Process and conduct diagnostic checks of the data before analysis

Remove unnecessary variables and participants who do not meet your inclusion criteria, since large datasets take longer to analyze. Conduct diagnostic checks to familiarize yourself with the nature and structure of your data and to find possible errors. Publications that use the data may provide some guidance.

Basic checks include:

- Confirm variable values and labels align with the codebook.
- Check for outliers and impossible values (e.g., scoring an 11 on a 7- point scale).
- Check the frequency of missing data.
- Confirm the unit of analysis (e.g., claim-level, person-level or state-level).
- Check the distribution of values for each variable — non-normal distributions such as skewed or bimodal distributions require special analytic techniques.

II. Secondary Data Resources

Tables 1 and 2 provide lists of publicly available and restricted datasets, respectively, in alphabetical order. For each dataset, the tables contain non-exhaustive summaries of their contents, high-level uses and limitations and example studies that use the data.

There are many secondary datasets not listed in this document. For example, ICPSR at the University of Michigan is a major repository of secondary datasets that allows users to filter by geography, restrictions, data format, methods and other characteristics.

IQVIA is a data vendor that licenses patient- and provider-level datasets spanning national and international health care settings. Check citations on your topic if you do not see a suitable dataset in these tables.

Table 1. Publicly Accessible Datasets

Dataset	Type and Contents	Good for	Limitations	Example Research
All of Us database from the National Institutes of Health	De-identified electronic health records, patient surveys, genome and wearable data.	Health outcomes research; genetic research; longitudinal analysis.	Participation in data sharing is voluntary, not representative. Researchers must register for data use.	Davis, M. R., Johnson, A. J., Denfeld, Q. E., Purnell, J., & Shannon, J. (2026). The utility of using the All of Us Research Program to examine AHA's Life's Essential 8. *Nursing Research*, *75*(1), 75-79.
American Community Survey (ACS)	Annual Census survey of U.S. households' demographic, social, economic and housing characteristics.	Estimates of NP supply, socioeconomic demographics, migration and location.	Before 2010 did not uniquely identify NPs; limited health variables; estimates for small areas not available unless using 3- or 5-year samples.	DePriest, K., D'Aoust, R., Samuel, L., Commodore-Mensah, Y., Hanson, G., & Slade, E. P. (2020). Nurse practitioners' workforce outcomes under implementation of full practice authority. *Nursing Outlook*, *68*(4), 459-467
Area Health Resource Files (AHRF)	Compilation of county-, state-, and national-level workforce and health system indicators.	Geographic comparisons; shortage area analysis; policy/environment covariates; workforce trends.	Area-level, not individual-level; some variables are several years older than the versions available in the source data.	McMichael, B. J. (2018). Beyond physicians: The effect of licensing and liability laws on the supply of nurse practitioners and physician assistants. *Journal of Empirical Legal Studies*, *15*(4), 732-71.
Consumer Assessment of Healthcare Providers & Systems (CAHPS)	Includes several Medicare and Medicaid patient and family experience surveys.	Analyses of patient experiences across different providers, health care settings, health agencies and health plans.	Large number of datasets that can be difficult to navigate; limited provider data; no clinical outcomes.	Kippenbrock, T., Emory, J., Lee, P., Odell, E., Buron, B., & Morrison, B. (2019). A national survey of nurse practitioners' patient satisfaction outcomes. *Nursing Outlook*, *67*(6), 707-712.
Health Insurance Exchange Public Use Files (Exchange PUFs)	Collection of 12 files with plan-level and issuer (company)-level information related to qualified health plans.	Analyses of insurance benefits, coverage and geographic availability and enrollee survey results.	No patient-level or provider-level data; some limits on plan details.	Dorilas, E., Hill, S. C., & Pesko, M. F. (2022). Tobacco surcharges associated with reduced ACA Marketplace enrollment: Study examines the impact of tobacco surcharges on enrollment in Marketplace health insurance plans. *Health Affairs*, *41*(3), 398-405.

Dataset	Type and Contents	Good for	Limitations	Example Research
Medical Expenditure Panel Survey (MEPS)	Large-scale survey of families/individuals, providers and employers with data on health care use, expenditures, sources of payment, insurance and quality.	Longitudinal analyses of insurance coverage and use, patient access to care, health care expenditures and source of payment.	No provider location in standard public-use files; geography is limited and more detailed information may require restricted access.	Li, Y., & Jones, C. B. (2021). Care received by patients from nurse practitioners and physicians in US primary care settings. *Nursing Outlook*, *69*(5), 826-835.
Medicare Provider Data Catalog	Searchable collection of Medicare datasets on services, fees, providers and health care settings.	Analyses of workforce trends, supply and cost of services across provider types and health care settings.	May require linking multiple files together.	Lee, K. A., O'Reilly-Jacob, M., Nguyen, T., Marriott, D., Costa, D. K., Weiss, M., & Yakusheva, O. (2025). Medicare Part B reimbursement and service volume differences between ambulatory nurse practitioners and physicians. *Nursing Outlook*, *73*(6), 102523.
National HIV Surveillance and Monitoring Data	CDC-funded data on HIV testing, prevention and services.	Analyses of health outcomes, policy impact, utilization of HIV services and preventive care.	HIV-specific and CDC-funded programs only. Continued data collection and government funding uncertain.	Weiser, J., Tie, Y., Crim, S. M., Riedel, D. J., Shouse, R. L., & Dasgupta, S. (2024). Do HIV care outcomes differ by provider type? JAIDS *Journal of Acquired Immune Deficiency Syndromes*, *96*(2), 180-189.
National Hospital Ambulatory Medical Care Survey	Hospital emergency and outpatient department services data.	Analyses of hospital discharges and emergency department visits, inpatient data, mortality and post-discharge outcomes, health services disparities.	Survey ended in 2022. Limited to participating hospital emergency and outpatient departments. Replaced by the National Hospital Care Survey (next row).	Wu, F., & Darracq, M. A. (2020). Physician assistant and nurse practitioner utilization in US emergency departments, 2010 to 2017. *The American Journal of Emergency Medicine*, *38*(10), 2060-2064.
National Hospital Care Survey	Nationally representative hospital-based data.	Similar to National Hospital Ambulatory Medical Care Survey.	Most recent data is from 2021; restricted data files are available for a fee.	Peters, Z. J., Ashman, J. J., Schwartzman, A., & DeFrances, C. J. (2022).National Hospital Care Survey demonstration projects: Examination of inpatient hospitalization and risk of mortality among patients diagnosed with pneumonia. *National Health Statistics Reports* No. 167.

Dataset	Type and Contents	Good for	Limitations	Example Research
National Plan and Provider Enumeration System (NPPES)	Centers for Medicare and Medicaid Services provider registry with national provider identifiers (NPIs), specialty and practice location. Updated monthly.	Workforce counts; mapping provider locations; analyses of workforce demographics, provider type, specialty.	Information may be out of date; self-reported taxonomies do not always reflect clinical role.	Xue, Y., Cai, X., & Poghosyan, L. (2023). Geriatric nurse practitioner supply and state scope-of-practice laws. *Journal of Nursing Regulation*, *14*(3), 4-13
National Practitioner Data Bank	Repository of medical malpractice payments and certain adverse actions involving health care practitioners.	Malpractice/adverse action analyses by provider type and state; disciplinary trends; state or policy comparisons.	Not a provider registry; public-use data are de-identified; reporting lag.	Dillon, D. (2024). Do transition to practice hour requirements make a difference in adverse action and medical malpractice payment reports: An analysis from the National Practitioner Data Bank. *Journal of the American Association of Nurse Practitioners*, *37*(6), 327.
National Sample Survey of Registered Nurses (NSSRN)	National workforce survey of nurses and advanced practice nurses that occurs about every four years.	Analyses of workforce demographics, education, earnings, hours, employment setting and burnout; change across survey waves.	The survey is not annual and has a gap from 2008-2018; limited geography information.	Markowitz, S., & Adams, E. K. (2022). The effects of state scope of practice laws on the labor supply of advanced practice registered nurses. *American Journal of Health Economics*, *8*(1), 66-98
Occupational Employment and Wage Statistics (OEWS)	Employment and wage data are aggregated at the national, state or metropolitan level by occupation.	Employment trend analyses; salary comparisons; assessing geographic variation in NP employment.	Only includes information for employed occupations; does not include patient claims, utilization or outcomes data.	Spletzer, J. R., & Handwerker, E. W. (2014). Measuring the distribution of wages in the United States from 1996 through 2010 using the Occupational Employment Survey. *Monthly Lab*. Rev., 137, 1.
Medicaid & Children's Health Insurance Program (CHIP) Open Data	State-level aggregated Medicaid and CHIP data such as enrollment counts and quality measures, National Average Drug Acquisition Cost dataset.	Policy analyses of drug costs and payment; enrollment; state health care quality.	Not individual-level; limited NP identifiers. Most peer-reviewed Medicaid research relies on restricted files designed for research use (see Table 2).	Three Axis Advisors. (n.d.). Issue brief: Section 206 of the bipartisan Prescription Drug Pricing Reduction Act (PDPRA), and similar proposals, seek to mandate responses by pharmacies to CMS' National Average Drug Acquisition Cost (NADAC) survey. Issue Brief.

Table 2. Selected Restricted-Access and Licensed Data Sources

Note: These sources may require data use agreements, fees, credentials and/or analysis within a specific computing environment.

Dataset	Type and Contents	Good for	Limitations	Example Research
Chronic Conditions Data Warehouse (CCW)	Medicare and Medicaid beneficiary, claims and assessment data linked by beneficiary with standardized variables and flags for chronic conditions. In addition, CCW provides underlying Medicare claims and enrollment data.	Longitudinal utilization and cost analyses in Medicare and Medicaid populations; construction of patient groups for analysis (e.g., by chronic condition).	Provider identifiers and location detail are limited and vary by file, restricting provider-level and geographic analyses.	Poghosyan, L., Liu, J., Perloff, J.,D' Aunno, T., Cato, K. D., Friedberg, M. W., & Martsolf, G. (2022). Primary care nurse practitioner work environments and hospitalizations and ED use among chronically ill Medicare beneficiaries. *Medical Care*, *60*(7), 496-503.
Healthcare Cost and Utilization Project (HCUP)	Family of hospital and emergency department discharge datasets with encounter-level information.	Inpatient, emergency and ambulatory department trends; analyses of hospital-based outcomes, costs and utilization; potentially longitudinal data.	No provider identifiers; linking across datasets is complex and may require additional files or permissions.	Kang, B., Fernando, T., Pang, J., Shirey, P., & Armstrong, D. P. (2025). Utilizing federal data sources to support nursing workforce analysis. *Policy, Politics, & Nursing Practice*, *26*(2), 97-109.
Merative MarketScan Research	Employer-sourced privately insured claims databases with longitudinal enrollment, utilization and cost data.	Longitudinal claims-based outcomes research; utilization, cost and employer-sponsored population studies.	Limited generalizability to non-employed populations. Provider identifiers available but may be incomplete or inconsistent across files.	Nugent, S. B., Lavin, R. P., Lee, J., Horn, B. P., & Damron, B. I. H. (2025). An assessment of nurse practitioner low-value care use in primary care. *The American Journal of Managed Care*, *31*(10), 294-298.

Dataset	Type and Contents	Good for	Limitations	Example Research
Military Health System (MHS) Data Repository	Comprehensive claims, enrollment, pharmacy, laboratory, radiology, clinical note and referral data from a single military health system.	Research on utilization of health care, costs and health outcomes focused on military members and their beneficiaries.	Direct access limited to Department of Defense personnel or sponsored and approved projects. Limited access available through Veterans Affairs (VA).	Richard, P., Gedeon, D., Yoon, J., Gibson, N., Narcisse, M. R., McCants, K., & DeGraba, T. (2025). Provider differences in costs, utilization and quality of primary care for traumatic brain injury in the military. *Value in Health*, *28*(10), 1506-1516.
Optum Claims Data	De-identified national claims database with linked pharmacy and lab data.	Longitudinal claims analyses including utilization, cost, pharmacy and lab data.	Includes only insured populations within contributing plans; may not be nationally representative. Provider identifiers available but may lack consistent specialty or role classification.	Tapper, E. B., Hao, S., Lin, M., Mafi, J. N., McCurdy, H., Parikh, N. D., & Lok, A. S. (2020). The quality and outcomes of care provided to patients with cirrhosis by advanced practice providers. *Hepatology*, *71*(1), 225-234.
T-MSIS Analytical Files (TAF) Research Identifiable Files (RIF)	Detailed Medicaid/CHIP claims not restructured for analysis within the CCW (top row, pg. 7).	Medicaid and CHIP encounter-level analyses; utilization, cost and outcomes in populations with lower incomes.	Strict data use agreements; substantial variation in data quality and completeness across states. Provider identifiers available but vary in completeness and reliability by state.	Luo, Q., Bodas, M., Vichare, A., Montellano, J., Jennings, N., Erikson, C., & Chen, C. (2023). Primary care provider Medicaid participation across the United States, 2016. *Journal of Health Care for the Poor and Underserved*, *34*(2), 703-718.
Veteran Affairs Corporate Data Warehouse	Veteran-level electronic health record data on patient and provider encounters.	Research on health outcomes and policy questions; utilization of services; longitudinal questions on veterans.	External researchers allowed with VA affiliation. Population skews older and male.	Rajan, S. S., Akeroyd, J. M., Ahmed, S. T., Ramsey, D. J., Ballantyne, C. M., Petersen, L. A., & Virani, S.S. (2021). Health care costs associated with primary care physicians versus nurse practitioners and physician assistants. *Journal of the American Association of Nurse Practitioners*, *33*(11), 967-974.