Data Source	Description	Strengths	Limitations[1]	Access
Centricity[2]	• GE Centricity is an electronic health record system that enables ambulatory care physicians and clinical staff to document patient encounters and exchange clinical data with other providers and information systems. • The database includes complete healthcare data, including prescription drugs, doctor visits, procedures, lab tests and results	• Recent data (lag time is less than 6 months)	• No cost data is available • Not nationally representative, may be difficult to generalize results • Pediatric and elderly populations may be underrepresented • Race is available, but incomplete	• A DUA is required • Data acquisition cost
CCW 5%[3] Sample with Part D events and Minimum Data Set (MDS)	• Chronic Condition Warehouse data with Part D outpatient claims and Minimum Data Set (MDS) assessments is populated with CMS data for a random 5% sample of ever enrolled Medicare beneficiaries eligible for coverage during a specified time period. (CY 1999 – 2009) • Available data: enrollment, Chronic condition flags, and claims including Inpatient, Carrier, Outpatient, Home Health Agency, Hospice, and Durable Medical Equipment. • Part D plan characteristics and outpatient claims are available starting in 2006 • MDS assessments are comprehensive survey mandated by the Centers for Medicare and Medicaid to be completed at defined intervals throughout a nursing facility stay	• Data lag time is about 3 years • Contains all Part D events, regardless of whether the beneficiary was enrolled in a managed care plan or a stand-alone prescription drug plan • Contains all MDS assessments, regardless of whether the beneficiary was enrolled in a managed care or fee-for-service plan • Large and representative sample of Medicare enrollees • Includes Part D plan descriptions	• Excludes health care data for individuals enrolled in Medicare HMOs or employer-based plans • Excludes children and adults less than 65 years old unless disable and eligible for Medicare	• A DUA is required • Data acquisition cost
i3 Invision™ Data Mart	• The i3 Innovus database includes enrollees of commercial insurance (primarily UnitedHealth). • Available data: inpatient and outpatient physician and facility claims, outpatient pharmacy claims, enrollment, lab test results and socioeconomic elements (e.g., income, education, race and ethnicity, life stage ad lifestyle indicators)	• Data lag time 90-day lag time • About 15 million covered lives per year • Available since May 2000	• Race is not available from claims data • Lab results data are only available for a subset of enrollees and tests • No plan payment data were available; only charges and patient's responsible portion of the payment were	• A DUA is required • Data acquisition cost

Data Source	Description	Strengths	Limitations[1]	Access
			available	
MarketScan Commercial Claims and Encounters; Medicare Supplemental and Coordination of Benefits (COB)	• The MarketScan Commercial Claims and Encounters (commercial) database provides patient level enrollment and claims data for the employees, their spouses and dependents who are insured by large employers. • The MarketScan Supplemental and COB (Medicare COB) database provides similar data for retirees with Medicare supplemental insurance paid for by large employers. • Available data: service level claims for inpatient and outpatient services, outpatient pharmacy claims (including retail, mail order and specialty drugs), and enrollment	• Recent data • Large number of covered lives • Complete payment/charge information, including amount of patient responsibility (both commercial and Medicare COB) and the Medicare covered payment (Medicare COB only) • Complete outpatient prescription drug data including the elderly	• Race is not available • Convenient samples from large employers; small and medium firms are not included • Population is generally healthier than the overall population	• A DUA is required • Data acquisition cost
Maryland Medicaid	• The Maryland Medicaid data provide eligibility and claims data for beneficiaries enrolled in FFS and MCO in the state of Maryland. • Available data include eligibility, inpatient and outpatient medical and outpatient pharmacy claims	• Availability of FFS claims and MCO encounters for enrollees • Race is available	• May have incomplete diagnosis and procedure codes in medical claims for enrollees who are dually eligible for Medicare and Medicaid	• A MOU is required[4] • Approval by the Department of Health and Mental Hygiene IRB is required • Data acquisition cost

Medicare Current Beneficiary Survey (MCBS)	• The MCBS is a continuous, multipurpose survey of a nationally representative sample of aged, disabled, and institutionalized Medicare beneficiaries. Two set of files are available: Access to Care (ATC) and Cost and Use (CAU).
	• The ATC file contains information on beneficiaries' access to health care, satisfaction with care, and usual source of care. It contains results from a supplement gauging beneficiaries' sources of information about Medicare and from a supplement surveying Medicare HMO members. Medicare FFS claims data for
		• The linked survey and Medicare claims data allows a broader range of analysis than would be possible using either survey or claims data alone
		• Race is available
		• Drug use is obtained through quarterly interviews
		• Includes a discreet but small population of institutionalized Medicare
			• Limit to Medicare beneficiaries (primarily people age 65 and older)
			  • CAU (CY 1992-2008)
			  • ATC (CY 1991-2009)
			• Drug records prior to 2006 need additional work to standardize names, dosage forms, and to categorize drugs into useful therapeutic classes
				• A Limited Data Set[5] (LDS) DUA is required;
				• A New Use Request is required for new data acquisition and a Reuse Request is necessary to reuse existing MCBS data for a new study[6]
				• Cost for data: $600 for each year and set of files (including Medicare claims)

Data Source	Description	Strengths	Limitations[1]	Access
	the beneficiaries are available. • The CAU survey includes information on the use and cost of all types of medical services, as well as information on supplementary health insurance, living arrangements, income, health status, and physical functioning. The linked Medicare claims provide use and cost information on inpatient hospitalizations, outpatient hospital care, physician services, home health care, durable medical equipment, skilled nursing home services, hospice care, and other medical services.	beneficiaries		
Medicaid Analytic eXtract (MAX)[7]	• The MAX files provide enrollment, service utilization and payment data for each State's Medicaid beneficiaries on a calendar year basis. MAX data are extracted from the Medicaid Statistical Information System (MSIS). • The data sets contain five separate files: Person Summary, Prescription Drugs, and Inpatient Hospital, Long Term Care, and Other Services files. • Data available CY 1999-2007	• Data are available for all 50 states and DC • Include beneficiaries in FFS and MCO programs but varies from state to state. • Data have undergone edit checks • Claims represent final action events • Race is available • Standardized data elements and formats within the year	• Data Lag time about 4 years • MCO encounter data are considered incomplete • Children enrolled in S-CHIP programs are often missing	• Requires approval of a DUA from CMS • Cost for data extraction • See request for Research Identifiable File[8] for details
National Veterans Health Administration (VHA)	• The National VHA data include standardized data submitted from all the VA facilities. • Multiple files are currently available: inpatient stay (acute, extended, observation and non-VA stay paid by VA), outpatient (visit and events), pharmacy (PBM), DSS National Laboratory Data (laboratory, laboratory results, radiology, pharmacy) and vital status • Future: linked CMS data (e.g., Medicare,	• Capture data for services rendered in any VA system • Availability of selected laboratory results • Availability of inpatient and outpatient pharmacy data • Recent data (data are updated through 2008)	• The majority of patients are men • Do not have data if the person is treated outside of the VHA system (except for care paid for by VA) • Require substantial data manipulation • No cost data in inpatient stay or outpatient encounters	• Requires VA IRB approval • Only authorized researchers can obtain data • All study team members have to be regular VA employees or are work without compensation staff • All work have to be performed within the VA firewall • No data acquisition cost

Data Source	Description	Strengths	Limitations[1]	Access
	USRDS, MCBS) for the veterans		• Incomplete race data	
Pennsylvania Medicaid	• The Pennsylvania Medicaid data provide eligibility and claims data for beneficiaries enrolled in FFS and MCO in the state of Maryland. • Available data include eligibility, inpatient and outpatient medical and outpatient pharmacy claims	• Availability of FFS claims and MCO encounters for enrollees • Race is available • Data is more recent than obtaining it from CMS	• May have incomplete diagnosis and procedure codes in medical claims for enrollees who are dually eligible for Medicare and Medicaid. • Data for MCO enrollees may be incomplete	• Data was obtained through a service contract to conduct retrospective drug use review • Current MOU/DUA/IRB requirements unknown • Data acquisition cost unknown
IMS LifeLink[TM9] PharMetrics	• This database includes data from 60 million unique members from >90 commercial health plans across the US. • Some Medicare beneficiaries are included • Available data: inpatient and outpatient claims, outpatient pharmacy claims (retail and mail order), enrollment	• Recent data (data are updated monthly with a 6-9 month lag) • Nationally represented sample • Well documented data dictionary • Charges and plan payment data are available	• Race is not available • Patient's responsible portion of payment (copay and deductible) is not present but can be imputed from allowed charge and plan payment • Enrollees older than 55 are under represented	• 10% sample is available for student dissertations and unfunded pilot projects without the cost DUA requirement • A DUA is required • Data acquisition cost
Philips VISICU eICU Research Institute (eRI) Data Warehouse	• This database is consists of medical records of admissions to intensive care units (ICU) that participate in the eICU program in both community and academic hospitals in the US. • Includes demographics, admission source, discharge location, admission diagnosis, past medical history, diagnosis, medications, flow sheet, notes, ventilation, laboratory, vital signs, and microbiology data	• Recent data • Totally de-identified data • Include a large number of admissions from 30 states as of January 2010 • Very rich data source	• Limit to ICU data only • Relatively small number of pediatric admissions • Most of the variables have text values • Cannot link a person across different health care system	• Data was obtained through a contract to provide analytic support
SEER-Medicare	• The SEER-Medicare data include data from the Surveillance, Epidemiology and End Results (SEER) registries and Medicare claims data for elderly people with cancer. • The SEER program provides registry data that includes clinical (e.g., staging and first course of treatment), demographic and survival information for cancer patients. • Medicare data include enrollment, and	• The availability of both the registry and Medicare claims data broadens the array of research that can be conducted than with either source alone in the area of diagnosis, treatment, survivorship, recurrences and terminal	• Limit to Medicare beneficiaries with cancer that resided in the SEER registry catchment area at the time of diagnosis • Do not have Medicare claims for HMO enrollees • Cannot identify specific chemotherapy agents	• Requires an approved DUA from NCI[10] • Data acquisition cost[11]

Data Source	Description	Strengths	Limitations[1]	Access
	claims from Medicare Provider Analysis and Review, Carrier, Outpatient, Home Health Agency, Hospice, Durable Medical Equipment and Part D outpatient prescription drug claims starting from 2007.	phase. • Medicare claims are available for a random sample of Medicare beneficiaries residing in the SEER areas who do not have cancer. This Information can be used in studies for comparative purposes. • Race is available from two sources.	administered as an inpatient • Limitations on Medicare claims also apply (e.g., non covered service) • Lag time (cancer cases diagnosed through 2007 and claims through 2009) • Request of restricted variables require approval from all SEER registries which can delay data acquisition	
United States Renal Data System (USRDS)	• USRDS is a national system that contains information on the entire US population of end-stage renal disease (ESRD) patients with few exceptions. Because the disease is a Medicare-qualifying reason, most patients have Medicare Claims information available. • USRDS is funded by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and the Centers for Medicare and Medicaid Services (CMS). USRDS staff collaborates with the United Network for Organ Sharing (UNOS). • The information available on patients includes general demographics, residence history, payer history, treatment history, wait list status and detailed transplant information.	• Since Medicare covers most cases of ESRD, administrative claims information is available for the majority of patients. • Race and ethnicity are available. • Date of death is available. • Treatment modality, residence, wait list status, and payer history are captured over time.	• Since people can qualify for Medicare based on age or ESRD-status, those who qualify based on age and then develop ESRD will have claims available pre-dating their ESRD diagnosis. This will not be the case for those qualifying for Medicare based on ESRD-status alone – unless the person also qualifies for Medicare based on a disability. • Laboratory values are captured at only one point in time using the CMS Medical Evidence Report, which is completed by the renal provider for each new ESRD patient. • There is a 2-3 year lag for complete data.	• A DUA is required • Data acquisition cost[12]

CMS = Center for Medicare and Medicaid Services

DC = District of Columbia

DUA = Data use agreement

FFS = Fee for service

MCO = Managed care organization

MOU = Memorandum of Understanding

NCI = National Cancer Institute

[1]Limitations beyond the general limitations for administrative claims data

[2]https://www2.gehealthcare.com/portal/site/usen/menuitem.b399d8492e44a6765c09cbd58c829330/?vgnextoid=ae0f4fb9efff5210VgnVCM100000382b3903RCRD

[3]http://ccwdata.org

[4]http://www.dhmh.maryland.gov/mma/html/datarequest.htm

[5]http://www.resdac.umn.edu/Medicare/data_available.asp#lds

[6] http://www.resdac.umn.edu/MCBS/requesting_data.asp

[7] http://www.cms.hhs.gov/MedicaidDataSourcesGenInfo/07_MAXGeneralInformation.asp

[8]http://www.resdac.umn.edu/Medicare/requesting_data_NewUse.asp

[9]http://www.imshealth.com/portal/site/imshealth/menuitem.0103f29c72c419cd88f611019418c22a/?vgnextoid=bc42650204850210VgnVCM100000ed152ca2RCRD&vgnextfmt=default

[10]http://healthservices.cancer.gov/seermedicare/obtain/requests.html

[11]http://healthservices.cancer.gov/seermedicare/obtain/cost.html

[12]http://www.usrds.org/cd_price2.htm

Information included in this document is valid as of February 28, 2012