Centers for Medicare and Medicaid Services
Medicare Data and Stroke Research
Goldmine or Landmine?

Judith H. Lichtman, PhD; Erica C. Leifheit-Limson, PhD; Larry B. Goldstein, MD

STROKE, a major public health problem in the United States, is the country’s fourth leading cause of death and a leading cause of serious disability. There are an estimated 6.8 million stroke survivors in the United States, with $795,000 people having a new or recurrent stroke annually. Its high prevalence and the functional deficits resulting from stroke impose a large burden on patients, their families, the healthcare system, and society. This is especially true among the elderly, for whom cerebrovascular diseases (stroke and transient ischemic attacks) are the second leading cause of hospitalization. Of those hospitalized with stroke, two thirds are aged ≥65 years; half are ≥70 years. With the aging of the population, the absolute number of elderly people with stroke is expected to increase in the coming decades.

The availability of national Medicare data relevant to the care and outcomes of older patients with stroke in the United States represents a valuable, yet largely untapped, resource. Medicare is the largest health insurance program in the United States, providing coverage for hospital, skilled nursing facility, home health, and hospice care, as well as outpatient services and prescription drugs, to millions nationwide. Approximately 52.3 million (~1 in 6) Americans were enrolled in Medicare during 2013, including 43.5 million of those aged ≥65 years; by 2030, the program is expected to serve >81 million people.

Stroke is one of the 10 highest contributors to Medicare costs. The estimated direct medical cost of stroke in the United States in 2012 was $71.6 billion, with an additional $33.7 billion in indirect costs attributable to loss of productivity (2010 dollars); these estimates are projected to triple by 2030.

There is a growing need for evidence-based knowledge about organizational strategies, structural characteristics, and healthcare policies that promote better care and outcomes for patients with stroke. Understanding patterns of care and outcomes are increasingly important because there is a national focus on quality of care, performance measures, costs, and efficiency. The purpose of this topical review is to provide an overview on the use of Medicare data for stroke research. We will highlight the strengths of the data and how they may be used as a cost-effective tool to enhance surveillance and outcomes research on elderly patients with stroke. We will also identify potential challenges in analyzing and interpreting Medicare data.

Why Use Medicare Claims Data for Stroke Research?

Randomized clinical trials, considered the gold standard for assessing treatment effects on patient outcomes, generally use strict inclusion/exclusion criteria that may limit their generalizability to the population as a whole. Moreover, clinical trials may not be ethical or feasible for the study of certain exposures on outcomes (eg, the study of air pollution on cardiovascular events). Community-based studies and registries that use primary data collection provide invaluable information on disease cause, care, and outcomes, but they can be costly, time consuming, and potentially challenging to conduct in diverse populations and regions. Medicare claims data are nationally representative, collected on an ongoing basis, and readily available in a timely matter.

How Are Medicare Data Collected, What Data Are Available, and How Can Researchers Gain Access to these Data?

Medicare data are based on Centers for Medicare and Medicaid Services (CMS) billing and administrative records and include a range of information about enrollees and their healthcare service usage, including demographic characteristics, coverage information, service dates, and International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) diagnosis and procedure codes. Researchers may request data files for a variety of care settings (eg, inpatient, outpatient, and skilled nursing facility), which can be linked by beneficiary to create an episode of care (see Figure 1 for an overview of available Medicare analytic files). Research identifiable files contain the most complete claims data, with directly identifiable, individual-level data on Medicare beneficiaries and providers. Access to these files requires application through the Research Data Assistance Center, a CMS contractor that...
provides assistance to academic and nonprofit researchers interested in using CMS data. These data requests also require a data usage agreement and CMS privacy board review to ensure privacy protection of beneficiaries and a justifiable need for identifiable data. Limited data sets (containing individual-health information but no direct identifiers) and public-use files/non-identifiable data (containing aggregated data and no identifiers) are also available for research; they do not require privacy board review and data requests can generally be submitted directly to CMS (Figure 2). Additional details on available Medicare data files and the data request process can be obtained from the Research Data Assistance Center.

### Who Is Covered by Medicare?

Medicare is the primary health insurer for those aged ≥65 years, <65 years with certain disabilities, and all ages with end-stage renal disease. Nearly all beneficiaries receive Part A hospital insurance benefits, including coverage for inpatient care, skilled nursing facilities, postinstitutional home care, and hospice services. Most beneficiaries (>91%) also choose to enroll in the Part B supplementary program, which provides coverage for physician services, outpatient care, durable medical equipment, and certain home healthcare services. In

### Research Identifiable Files (RIF)
- Beneficiary-level files containing identifiable information and protected health indicators
- Requires Data Use Agreement with CMS
- Reviewed by CMS Privacy Board to ensure protection of patient privacy and justifiable need for identifiable data

### Limited Data Set (LDS)
- Beneficiary-level files containing protected health indicators
- Many of the identifiable fields are encrypted, blank, or ranged
- Requires Data Use Agreement with CMS, but not Privacy Board approval

### Public Use Files (PUF)/Non-Identifiable Files
- Files containing either aggregated data or non-identifiable beneficiary-level information
- Available in the public domain

### Figure 1. Available Medicare claims data files.
CMS indicates Centers for Medicare and Medicaid Services; DME, durable medical equipment; FFS, fee-for-service; HCPCS, Healthcare Common Procedure Coding System; ICD-9-CM, International Classification of Diseases, Ninth Revision, Clinical Modification; NDC, National Drug Code; NPI, National Provider Identifier; SNF, skilled nursing facility; and UPIN, Unique Physician Identification Number.

### Figure 2. Three types of Medicare data according to privacy level. CMS indicates Centers for Medicare and Medicaid Services; LDS, limited data set; PUF, public use files; and RIF, research identifiable files.
addition, >70% of beneficiaries purchase prescription drug benefits under the Part D program. The majority of beneficiaries (≈72%) receive Medicare coverage through the traditional fee-for-service program, with the remainder electing to enroll in private health insurance plans under Medicare Advantage (Part C). These private managed care plans provide Part A and Part B services and receive capitated payments for enrolled beneficiaries (ie, payment per beneficiary rather than payment per service provided). Part C enrollees are generally excluded from analyses because detailed information on their use of services is often unavailable in the data.10

What Are Some of the Advantages and Uses for Medicare Data for Stroke Research?

There are several advantages associated with using Medicare data for stroke-related research (Figure 3). Many domains of care are included, and depending on specific research questions, data can be used to examine a broad range of outcomes, including in-hospital outcomes, discharge destinations, short- and long-term mortality, readmissions, and the use of outpatient Medicare resources across a variety of care settings (hospital outpatient, home health agency, hospice, and skilled nursing facility). Research files for CMS data include a unique, encrypted beneficiary identifier that is consistent across the different Medicare file types, permitting researchers to link and analyze information across the continuum of care. Because the data are continuously collected as part of the Medicare billing system, they include every patient encounter associated with a billing record. The data are also available from a single source and are provided in standard formats.

Medicare data can provide a complete national sample with ample power to conduct subgroup analyses, such as by patient demographic characteristics or geographic region. This is particularly valuable because of the lack of a dedicated national stroke surveillance system in the United States to track national trends in stroke hospitalizations and outcomes, as well as provide insights into patterns of stroke care.11 The data allow for stroke surveillance at the patient, hospital, community, state, and national levels. For example, the Centers for Disease Control and Prevention, in collaboration with CMS, published an atlas showing geographic patterns in age-adjusted stroke hospitalization rates by county that demonstrated variations across regions, with higher rates in the southeastern Atlantic region, often referred to as the stroke belt.12 Medicare data have also been used to track poststroke events, such as recurrent stroke hospitalizations. A study using patient-linked data including >2.5 million beneficiaries found heterogeneity in the occurrence of recurrent ischemic stroke by geographic region from 1994 to 2002.13 Overall, recurrent stroke rates declined by ≈5% over the time period, but temporal patterns of patient-level outcomes varied markedly by region; recurrence rates decreased within sections of the Southeast stroke belt and increased in counties in the Midwestern and Western regions of the United States. A more recent study using 2006 Medicare fee-for-service data assessed national rates and variations in potentially preventable readmissions after stroke across hospitals and regions (Figure 4).14 Among 307 887 ischemic stroke discharges, 14.4% of patients were readmitted within 30 days, and 11.9% of these readmissions were the result of a preventable cause (eg, pneumonia). There was regional variation in preventable readmissions, with the highest rates observed in the Southeast, Mid-Atlantic, and United States territories and the lowest in the Mountain and Pacific regions, as well as considerable hospital-level variation both within and between regions (Figure 4).

Provider characteristics, such as operator expertise and infrastructure (eg, hospital size), can influence receipt of care and patient outcomes. Medicare data provide an opportunity to identify treatment variations over time and across providers and geographic locations, and the data can be used to help benchmark the quality of healthcare delivery for stroke-related therapies and interventional procedures. For example, a nationwide study of recombinant tissue-type plasminogen activator use among Medicare beneficiaries hospitalized for ischemic
stroke from 2005 to 2007 identified 11,884 cases of recombinant tissue-type plasminogen activator use (2.4% treatment rate), with wide variations in treatment rates by hospital (of 4750 hospitals, 64.2% recorded no treatment with recombinant tissue-type plasminogen activator, 18.8% had rates from 0% to 2.4%, and 0.9% had rates >10%). Smaller bed size, lower population density, and location in either the South or the Midwest were each associated with lower recombinant tissue-type plasminogen activator use. Another study using national Medicare fee-for-service data from 2005 to 2007 found higher 30-day mortality rates after carotid artery stenting for patients treated by low- versus high-volume operators (odds ratio, 1.9; 95% confidence interval, 1.4–2.7) and patients treated by operators early in their careers versus later (odds ratio, 1.7; 95% confidence interval, 1.2–2.4) among a sample of 24,701 patients who underwent the procedure by 2339 operators.

Hospital certification programs have been embraced as one means of optimizing stroke quality of care and outcomes, but national data assessing their effect on outcomes are limited. Because Medicare data can be aggregated to different units of analyses (eg, patient, hospital, and region), their use provides the opportunity to assess organizational aspects of care and outcomes. Two studies using national Medicare data from 2006 found that Joint Commission Primary Stroke Center–certified hospitals had lower 30-day risk-standardized mortality rates for ischemic and hemorrhagic stroke when compared with noncertified hospitals, but there was no difference in 30-day readmissions. Medicare data also provide service use information that can be used to assess the care and outcomes for patients with stroke treated in small, rural hospitals—a group of care institutions not routinely captured in registries and cohort studies.

Medicare data can provide insights into regional variations in spending for stroke care. Among 327,521 patients with ischemic stroke identified from 2008 to 2009 using Medicare claims data, there was considerable regional variation in the use of high-risk computed tomography (24 head scans in 1 year) and risk- and price-adjusted expenditures per patient with stroke (ranging from $31,175 in Salem, MA, to $61,895 in McAllen, TX) during the first year after the index stroke event.

Finally, Medicare data may be linked with other databases and registries, either directly by using high-quality identifiers (eg, Social Security Numbers) or indirectly, to provide an efficient and cost-effective method to enhance follow-up strategies and enrich the description of established cohorts. National data can be linked longitudinally to facilitate the tracking of beneficiaries over time and across facilities, states, and provided services, potentially reducing losses to follow-up (eg, because of participant relocation). Event dates are available, which add precision to identifying outcomes, such as mortality, recurrent events, use of resources, and related care costs. For example, the ascertainment of vital status is high for elderly fee-for-service beneficiaries. Linkage with Medicare data may reduce participant burden to recall details of healthcare use, allowing investigators to allocate resources to collect patient-centered outcomes not available in administrative data (eg, functional recovery and quality of life).

Several observational cohorts/registries relevant for stroke research have linked their participants with Medicare administrative data, including Get With The Guidelines-Stroke, Reasons for Geographic and Racial Differences in Stroke (REGARDS), Chicago Health and Aging Project, Women’s Health Initiative, and Health and Retirement Study, among...
others. For example, the Chicago Health and Aging Project has used Medicare data to identify strokes among its participants and assessed the relationships of cognitive functioning, neighborhood cohesion, and psychosocial distress with stroke risk and mortality. The Get With The Guidelines-Stroke registry has used linked Medicare data to assess outcomes beyond the index hospitalization, examining hospital-level and racial/ethnic differences in 30-day and 1-year mortality and rehospitalization. Linked Medicare data have also been used in conjunction with clinically rich registry/cohort data to conduct sensitivity analyses that fill important gaps in the understanding of the effect of missing clinical variables (ie, information not currently collected as part of administrative billing data) on outcomes in administrative claims-based analyses and to assess the validity of claims-based diagnoses. Analyses of 127,950 Medicare fee-for-service beneficiaries hospitalized at 782 hospitals participating in Get With The Guidelines-Stroke showed that adding stroke severity (currently unavailable in Medicare data) to an administrative claims model of 30-day mortality significantly improved model discrimination (C statistics 0.86 versus 0.77; P<0.001) and modified mortality performance rankings for a large proportion of hospitals. Both REGARDS and the Women’s Health Initiative have reviewed the validity of claims-based algorithms for stroke. Using physician-adjudicated events as the gold standard, these studies found that algorithms including primary discharge diagnosis codes for hemorrhagic or acute ischemic stroke (ICD-9-CM codes 430, 431, 433.x1, 434.x1, and 436) had high specificity and negativity predictive values (>99%), high positive predictive values (85%–93%), but lower sensitivities (59%–75%) for identification of stroke.

What Are the Potential Landmines of Using Medicare Data?

Despite the many potential uses of Medicare data for stroke research, there are also important limitations that should be highlighted (Figure 3). Because Medicare data are collected for payment purposes, they are limited to submitted claims and services covered by CMS and data quality is often best for conditions and procedures that affect payment. CMS does not pay 100% of covered services. Medicare beneficiaries often have cost sharing requirements, including deductibles and coinsurance. Claims are generated for these services, but it is not possible to determine the true out-of-pocket cost to the beneficiary or whether the noncovered costs were paid by supplemental insurance. Generalizability can be an issue because enrollment varies across the Medicare plans such as fee-for-service, Medicare Advantage, and Part D. One strategy to minimize this concern is restriction of the inclusion/exclusion criteria for specific analyses (eg, restrict to beneficiaries with continuous fee-for-service care for a specified preindex and postindex event timeframe; exclude individuals discharged against medical advice or enrolled in hospice care). In addition, many clinical details that could be important for stroke research are lacking in administrative billing data. Laboratory results, imaging reports, data reflecting stroke cause, cognitive and functional status, and stroke severity are not collected. The lack of stroke severity in administrative data, in particular, has been a focus of much recent debate because it is a strong, independent predictor of mortality outcomes and improves the discrimination of hospital-level models of short-term mortality. Its absence in administrative data has garnered increased attention in the stroke community as CMS plans for the 2016 implementation of 30-day risk-standardized ischemic stroke outcomes measures for public reporting linked with payment. A recent American Heart Association/American Stroke Association Scientific Statement highlighted the importance of including stroke severity in models used to compare hospital performance. This statement noted the need for further research evaluating the ability of models without recommended variables (particularly stroke severity) to produce discrimination in hospital performance that is comparable with models that include these variables and the need for more research to ascertain stroke severity using available resources, such as administrative billing codes or electronic health records. The development of models using Medicare data without stroke severity continues to be a point of discussion among researchers, clinicians, care providers, payers, and those involved with healthcare policy. In the future, including codes to capture initial stroke severity would be useful for both research and reimbursement purposes; however, there is no current mechanism for CMS to develop codes that may be of interest to researchers but are not directly related to billing codes for payment reimbursements.

The validity of claims-based algorithms for identifying stroke and other conditions is important when considering the usefulness of Medicare data for stroke research. For stroke, the accuracy of case identification can be optimized using ICD-9-CM codes shown to have good sensitivity and specificity based on physician-adjudicated validation studies. It is possible to refine the accuracy of identifying events further by excluding procedures that may represent complications that may have occurred during the hospitalization. A commonly used strategy to address this problem is to identify codes for comorbid conditions in the database before the index event. These conditions can be captured by individual ICD-9-CM codes and used to calculate comorbidity indexes, or they may be based on codes grouped into clinically coherent categories. Because Medicare data depend on diagnoses recorded by healthcare providers in medical records or coded by medical record/billing personnel in discharge abstracts and claims forms, variations in ICD-9-CM coding practices, unrelated to patient characteristics, across care settings or regions could introduce biases in risk-adjustment models or case ascertainment. Such variations in coding may occur as a result of differences in diagnostic practices (eg, tendency to order diagnostic tests or refer patients to specialists, interpretation of pathological and radiological findings) or possible upcoding, with conditions recorded on submitted claims forms in a way that increases reimbursement payments from CMS.

Finally, it is important to recognize that there can be analytic challenges in using administrative data, including data management and selection of analytic approaches for multilevel...
data (eg, statistical techniques that account for clustering of cases within hospitals). Given the potentially large number of observations involved in these types of analyses, care is also needed when interpreting study results and distinguishing meaningful differences from merely statistically significant ones.

Conclusions
The availability of national Medicare data relevant to the care and outcomes of patients with stroke represents a potential goldmine for researchers. Because there is a need for high-quality national data reflecting the occurrence of stroke and its outcomes in the United States, with an understanding of the potential landmines associated with its use, Medicare data can provide critical information to address important epidemiological, clinical, and policy questions. Developing collaborative relationships between CMS and investigators could further inform the development of billing codes that reflect varied aspects of clinical care that affect both costs and quality, as well as enhance the use of this rich data resource for stroke research.

Disclosures
None.

References


**KEY WORDS:** health services research, Medicare, methods, stroke
Centers for Medicare and Medicaid Services Medicare Data and Stroke Research: Goldmine or Landmine?
Judith H. Lichtman, Erica C. Leifheit-Limson and Larry B. Goldstein

Stroke. 2015;46:598-604; originally published online January 15, 2015;
doi: 10.1161/STROKEAHA.114.003255
Stroke is published by the American Heart Association, 7272 Greenville Avenue, Dallas, TX 75231
Copyright © 2015 American Heart Association, Inc. All rights reserved.
Print ISSN: 0039-2499. Online ISSN: 1524-4628

The online version of this article, along with updated information and services, is located on the World Wide Web at:
http://stroke.ahajournals.org/content/46/2/598

Permissions: Requests for permissions to reproduce figures, tables, or portions of articles originally published in Stroke can be obtained via RightsLink, a service of the Copyright Clearance Center, not the Editorial Office. Once the online version of the published article for which permission is being requested is located, click Request Permissions in the middle column of the Web page under Services. Further information about this process is available in the Permissions and Rights Question and Answer document.

Reprints: Information about reprints can be found online at:
http://www.lww.com/reprints

Subscriptions: Information about subscribing to Stroke is online at:
http://stroke.ahajournals.org//subscriptions/