Administrative and Claims Records as Sources of Health Care Cost Data

Gerald F. Riley, MSPH

Background: Many economic studies of disease require cost data at the person level to identify diagnosed cases and to capture the type and timing of specific services. One source of cost data is claims and other administrative records associated with health insurance programs and health care providers.

Objective: To describe and compare strengths and limitations of various administrative and claims databases.

Data and Methods: Data sources included claims and enrollment records from Medicare, Medicaid, and private insurers; Veterans' Health Administration records; state hospital discharge datasets; Healthcare Cost and Utilization Project hospital databases; managed care plan data systems; and provider cost reports. Claims provide information on payments, whereas cost reports yield resource costs incurred to produce services. Administrative data may be significantly augmented by linkage to disease registries and surveys.

Results: Administrative data are often available for large, enrolled populations, have detailed information on individual service use, and can be aggregated by service type, episode, and patient. Service use and costs can often be tracked longitudinally. Because they are not collected for research purposes, administrative data can be difficult to access and use. Limitations include generalizability, complexity, coverage and benefit restrictions, and lack of coverage continuity. Linked datasets permit identification of incident cases of disease, and analyses of health care costs by stage at diagnosis, phase of care, comorbidity status, income, and insurance status.

Conclusions: Administrative data are an essential source of information for studies of the financial burden of disease. Cost estimates can vary substantially by specific measures (payments, charges, cost to charge ratios) and across data sources.

Key Words: claims data, administrative data, health care costs

(*Med Care* 2009;47: S51–S55)

Health care accounts for a large and increasing share of the US economy.[1] The way this money is spent, and the distribution of these costs among payers have crucial implications for access, affordability, and outcomes of care. Analyses of the economic burden of cancer care are essential to the formulation of policies regarding health insurance, the allocation of resources, and mitigation of disparities in the receipt of health care services.

Many economic studies of disease require data at the person level to identify incident cases and to capture the type and timing of specific services for each case. It is often necessary to ascertain the costs associated with individual services, or costs incurred over a given time period. One source of cost data is claims and other administrative records associated with health insurance programs and health care providers. Such data are usually available at the person level and records are often kept of individual services received and the amounts paid for those services. Identifiers are sometimes available that permit the aggregation of costs over episodes of treatment, or by patient.

This article will review some common sources of administrative data, their strengths and limitations, and some applications. Differences among payment, charge, and cost report data will be explored. Some methods of augmenting administrative data, such as linkage to surveys and other datasets, will be discussed. Many examples in the article are drawn from Medicare databases, but most of the conclusions apply to other administrative data systems as well.

USING ADMINISTRATIVE DATA TO ESTIMATE COSTS

Sources

A common source of administrative data is health insurance claims and encounter data associated with Medicare, Medicaid, the Veterans' Health Administration (VHA), and other public programs. Private insurance claims datasets are also available through vendors such as MedStat and Mercer. Individual level data on utilization and costs is sometimes available from managed care plans, state hospital discharge datasets, hospitals, and other providers of services. The Healthcare Cost and Utilization database, maintained by the Agency for Healthcare Research and Quality, contains detailed data on hospital services from specific states.[2] An important feature of many administrative data systems is that they apply only to specific groups of individuals, and therefore have limited generalizability to the US population. Medicare has the broadest population-based administrative data system in the health arena, covering about 97% of the elderly. It should be noted that not all administrative record systems are available to researchers and those that are available usually have restrictions on use of the data.

From the Office of Research, Development, and Information, Centers for Medicare and Medicaid Services, Baltimore, Maryland.

Reprints: Gerald F. Riley, MSPH, Centers for Medicare and Medicaid Services, 7500 Security Blvd., Mail Stop C3-21-27, Baltimore, MD 21244. E-mail: gerald.riley@cms.hhs.gov.

ISSN: 0025-7079/09/4700-0051

Payments Versus Charges

For a given service, administrative records may contain data on billed charges and/or the amount that was actually paid for that service. Billed charges represent the "list price" for services established by the provider, and may represent the actual price paid by uninsured individuals and some private insurers. Payment amounts for insured patients are often established by the insurer independent of billed charges, or negotiated between the provider and insurer. Consequently, billed charges and actual payments often differ substantially.

Claim payment amounts represent 1 definition of cost, namely the transaction price paid for services. They do not necessarily reflect the value of resources that were used to produce the services. Different insurers often use different payment formulas, and the cost of a given service estimated from Medicare payments, for example, may differ significantly from the cost of the same service estimated from Medicaid payments. In recent years, Medicare has moved to prospective payment systems for many of its covered services. Prospective payments are generally based on input costs, but these input costs are averaged over many providers. Although some adjustments are made for individual provider circumstances, payment amounts are relatively standardized and do not reflect all the natural cost variation among providers. Before the shift to prospective payment systems, most Medicare institutional claim payment amounts were based on estimated input costs of individual providers.

Billed charges are sometimes used to estimate costs of services, particularly when payment data are not available. Charges may be related to the input costs required to provide a service, but frequently they are not. For example, providers may not have a clear idea of what it does cost to produce a service, or they may knowingly set some charges at a higher or lower amount than their production costs for business reasons. At the hospital level, charges are often much higher than the cost of producing services, as reflected in hospital cost reports.[3]

Charges and payments often differ significantly because most payers do not pay on the basis of charges, but on formulas like Diagnosis Related Groups or fee schedules. Under Medicare, payments tend to be much lower than charges, and the relationship between payments and charges varies by type of service.[4] For example, in 2006 Medicare payments (exclusive of beneficiary cost-sharing amounts) were only 27.9% of billed charges for inpatient hospital services and 34.3% of billed charges for physician/supplier services (Table 1). The relationship between payments and charges has also changed over time—payments as a percent of charges have generally declined, especially for outpatient hospital services.

TABLE 1. Medicare Payments as a Percent of Billed Charges, For Selected Services, 1991 to 2006

Type of Service	1991	1996	2001	2006
Inpatient hospital (short-stay)	50.0	51.4	38.9	27.9
Skilled nursing facility	NA	45.2	64.6	65.1
Home health	72.9	71.8	103.8	102.1
Hospice	NA	99.3	94.4	92.2
Physician/supplier	NA	42.2	40.2	34.3
Hospital outpatient	39.1	31.2	25.0	16.4

Data were for services provided in the fee-for-service sector. Medicare payments do not include beneficiary cost-sharing amounts. This table is based on data contained in the Health Care Financing Review Medicare and Medicaid Statistical Supplement, 2007, Tables 5.1, 6.1, 7.1, 8.1, 9.1, 10.1.
NA indicates not available.

Cost Reports and Cost to Charge Ratios

Some researchers use cost to charge ratios to estimate the cost of producing services.[5–6] This method is commonly applied to hospital services, where input costs are often available from hospital cost reports filed with the Centers for Medicare and Medicaid Services or with state regulatory agencies. Such reports are used in establishing payment amounts and provide detailed accounting information on direct and indirect costs by department, as well as data on utilization and facility characteristics. To estimate the cost of a given service, the charge for that service is multiplied by the ratio of costs to charges for the hospital that provided the service. Accuracy is better if cost to charge ratios are applied at the level of the hospital department, rather than a single hospital-wide cost to charge ratio.[5] Cost to charge ratios can also be used to estimate the costs of services provided by other institutional providers that file cost reports, such as skilled nursing facilities, home health agencies, hospices, and dialysis facilities. Using provider-specific cost reports permits researchers to measure variation among providers in the cost of producing services. A disadvantage of using cost reports is that they are more complex than claims records and the degree of difficulty involved in working with cost reports is much greater. The Healthcare Cost and Utilization database includes hospital-specific cost to charge ratios for some of their hospitals.[2]

Several technical issues related to the use of Medicare cost-to-charge ratios have been documented by the Research Data Assistance Center, which provides information and assistance to researchers who use Medicare and Medicaid data.[7] A significant issue is that the revenue centers listed on claims records are not the same as departments in the costs reports, and there is not a readily available crosswalk between the two. Cost-to-charge ratios sometimes cannot be computed at the department level, and a higher level of aggregation may be needed. Research Data Assistance Center also points out that some providers have multiple cost reports in a given year, and that some extreme values can be found in cost-to-charge ratios for various reasons.

Some administrative databases do not have payment or charge data at the individual service or patient level. An example is the VHA data system, which includes detailed utilization data at the patient level and cost reports at the facility level. Several methods have been developed to estimate the cost of services using VHA data, sometimes in combination with data from external sources.[8]

LINKING ADMINISTRATIVE DATA TO OTHER DATABASES

The accuracy of cost estimates can sometimes be improved by linking 2 or more datasets, particularly when individuals have more than 1 kind of insurance or program entitlement. For example, many low income elderly and

disabled people are dually eligible for Medicare and Medicaid coverage. Because Medicare does not pay for nursing home services, linking Medicaid to Medicare records can provide information on both acute and long-term care costs. Many Medicare beneficiaries also are entitled to veterans benefits, and before implementation of the Medicare drug benefit, many beneficiaries received drug benefits through the VHA. Linking claims for different types of services can also be useful, such as linking hospital and physician bills for the same stay, or grouping claims into a single episode of care.

There are some significant barriers to data linkage. Researchers cannot always get access to identifiers that enable linkage among administrative datasets, either because common identifiers are not collected, or they cannot be used due to privacy concerns. Matching of records can be technically difficult, especially if there are many inaccuracies or missing values in the data. Sometimes administrative datasets can be combined without linking records at the individual level. Studies of aggregate costs, for example, may combine summary data from 2 or more administrative data systems without record linkage at the person level.

Disease Registries

Linking administrative data to disease registries has significantly increased the ability of researchers to analyze the costs of treating specific conditions. Two well-known examples are the Surveillance, Epidemiology, and End Results (SEER)-Medicare linked database, and the US Renal Data System.[9–10] These linkages have made possible a wide range of studies on cancer and renal disease that could not have been done using any single dataset. A key advantage of such linked data is that incident cases of disease can be more accurately identified and study cohorts defined.[11] Claims can then be organized by dates of service to identify baseline costs and comorbidities, as well as utilization and costs of care following diagnosis. Costs can also be estimated for the various phases of care, and phase-specific costs can be combined to create estimates of lifetime medical care costs between diagnosis and death.[12] The linkage of Medicare claims with SEER data has also enabled researchers to estimate medical care costs specifically attributable to cancer.[13–14] Lastly, disease registries also supply important clinical detail such as stage at diagnosis.

It should be noted that linked disease registry and administrative data are not as current as administrative data alone. For example, the Centers for Medicare and Medicaid Services and the National Cancer Institute must work with over a dozen tumor registries to obtain identifying information on SEER cancer cases, and then match this information to Medicare enrollment records. Significant time is also required to extract Medicare claims files for each matched sample of cancer cases. Consequently, SEER-Medicare linked records are usually not available to researchers until 3 to 5 years after cases are diagnosed.

Surveys

The usefulness of administrative data can also be enhanced by linkage to surveys. Examples of surveys linked to Medicare data include the Medicare Current Beneficiary Survey, the National Health Interview Survey, the National Health and Nutrition Examination Survey, the Health and Retirement Study, and the National Long Term Care Survey. These linkages enable researchers to examine the association of health care costs with income, institutionalization, functional status, insurance status, and other factors.[15–17] Survey reports can also fill in information on use and costs of services that are not covered by Medicare or other insurers. This is especially important for services that are commonly paid for out-of-pocket, like nursing home and dental care. Costs can be related to patient attitudes toward care, program knowledge, and reported access problems.

The linkage of survey and administrative data often presents special challenges. Authorizations must be obtained from the agency maintaining the administrative data and from the organization conducting the survey. Prospective respondents may have to be informed about any link with administrative data, and sometimes explicit respondent permission is needed. The survey organization must also arrange to obtain identifying information on respondents to carry out data linkage. In addition, sample sizes are often small, which is a particular problem when dealing with cost data. Because cost data tend to have large variances and are highly skewed, any estimates based on small samples may not be very reliable. This is a particular problem in estimating costs associated with specific diseases, which may be uncommon in samples from general, multipurpose surveys.

It can be difficult to reconcile claims data with utilization reported by the respondent. Reconciliation is important to avoid double counting services that both appear in the claims and are reported by the respondent directly. For example, services recorded on Medicare claims are routinely matched to services reported by Medicare Current Beneficiary Survey respondents. For services appearing in the claims and the survey, data are combined from both sources to create a single utilization record for each service.[18]

STRENGTHS AND LIMITATIONS OF ADMINISTRATIVE DATA

Strengths

Administrative records have important strengths with respect to measuring the cost of care (Table 2). The data often apply to a well defined population and include subgroups such as the institutionalized and cognitively impaired that tend to be underrepresented in surveys. Populations covered by these datasets may be defined by enrollment or by use of specific providers. Data definitions and formatting are usually standardized and are sometimes well documented for the user. In claims databases, procedure codes and dates of service permit the researcher to aggregate costs by type of service and to identify discrete episodes of care. Another important advantage is that service use and costs can often be tracked longitudinally over multiple years. This permits the investigator to measure costs associated with separate domains of care, such as initial treatment, surveillance, and end of life care. Lastly, administrative data are less costly to acquire than primary data because they have already been collected for other purposes.

Medicare data have recently been improved through the development of the chronic condition warehouse (CCW), which combines several sources of Medicare data, and facil-

TABLE 2. Strengths and Limitations of Administrative Data For Estimating Health Care Costs

Strengths	
General	Standardized definitions and documentation
	Less costly to obtain
	Data can be linked to disease registries and surveys
Population	Large, well defined populations
	Include vulnerable subgroups
	Data can be aggregated by episode or person
	Longitudinal studies possible
Payment data	Usually good data on transaction costs
	Data available at the individual service level
	Cost impact of comorbidities can be estimated
	Data frequently reflect provider variation (cost reports, charges)
Limitations	
General	Data collected for non-research purposes
	Restrictions on use (privacy, proprietary)
Population	Limited generalizability to the entire US population
	Data on managed care enrollees frequently absent from insurance claims databases
	Lack of continuity of insurance or program coverage
	Difficult to identify incident cases of disease
	Patient identifiers may change over time, making longitudinal analyses more difficult
Benefits/coverage	Claims limited to covered services
	Claims are grouped by bill type, reflecting the benefit but not necessarily visits or episodes of care; an episode may involve multiple claims within and across bill types
	Coding of diagnoses and procedures more closely related to billing requirements than to medical records
	Claim payments subject to benefit design features
Payment data	Clinical data are very limited–underlying reason for service and outcome not available
	Claims may contain biases due to provider efforts to maximize payment
	Some data sources (eg, those based on provider accounting systems) may be difficult to use
	May not be suitable for estimating the cost of new procedures or services

itates their use by researchers.[19] The CCW is designed to support studies on improving care for chronically ill beneficiaries. Among other features, the CCW contains 21 annual chronic condition flags denoting the presence of specific diagnostic codes on Medicare claims.

Limitations

Because administrative data have been collected for other purposes, they are not necessarily in a format that is intelligible or convenient to researchers. Several limitations should be kept in mind.

Population

Claims data are sometimes not available for enrollees in managed care plans because such plans tend to be paid by insurers on a capitation basis. In February 2009, almost 10.8 million beneficiaries were enrolled in Medicare managed care and other capitated plans, representing about 24% of beneficiaries, and this percentage is growing. In some insurance systems such as Medicaid, people can move into and out of coverage, presenting problems of continuity for longitudinal studies.[20]

Benefits and Coverage

Claims data are limited to services covered by the insurer. Until the implementation of the Part D benefit, Medicare had very little data in its administrative system on prescription drug costs, and still has almost no information on nursing home use. Claim payment amounts are also a function of benefit design features. For example, data may not be available for services received after a benefit cap has been reached. Differences in benefits and payment methods create challenges in comparing costs across systems or insurers.

Identifying Chronic Conditions

It is difficult to distinguish incident from prevalent cases of disease based on claims data.[21] This complicates efforts to attribute costs to specific conditions and phases of treatment.

Interpreting Payment Data

Payments represent the amount a given insurer paid, but patient cost-sharing amounts will be in separate data fields on the claim record. Often patient cost-sharing represents a substantial portion of the total cost of a service. For example, in 2006 patient cost-sharing represented 21.5% of provider payments under Medicare Part B, which covers physician and outpatient services, medical supplies, and durable medical equipment (Table 3).[4] (Note that Medicare cost-sharing amounts are often paid by a supplemental insurer, such as Medicaid or a Medigap policy.) Cost sharing for certain services, such as mental health services, can be higher than for other services. Payment amounts listed on claim records may represent "interim payments,"

TABLE 3. Medicare Payments and Patient Liability for Covered Services, 2006

	Dollars (in Millions)	
Type of Liability	**Part A**	**Part B**
Total	$165,833	$164,077
Medicare payments	$151,917	$128,755
Patient liability*	$13,916	$35,322
Patient liability as percent of total	8.4%	21.5%

Data were for services provided in the fee-for-service sector. This table is based on data contained in the Health Care Financing Review Medicare and Medicaid Statistical Supplement, 2007, Tables 3.2 and 4.1.

*Patient liability consists of deductible, coinsurance, nursing facility coinsurance (Part A only), and balance billing (Part B only).

which are subject to subsequent reconciliation and adjustment. Payment systems and benefit packages may also change over time, which can affect longitudinal and trend analyses. For example, Medicare substantially revised its payment methodology for home health services pursuant to the Balanced Budget Act of 1997, and this was followed by a large decline in utilization and costs for home health care. Lastly, it may take time to develop payment amounts for new or experimental services. For example, a new hospital service may initially be included in an existing Diagnosis Related Groups category that does not accurately reflect the average cost of producing that service.

Claims data may contain biases due to efforts by providers to maximize payments. For example, upcoding, or "coding creep" may increase expenditures in some programs over time.[22–23] Expenditures may also increase as providers improve the accuracy and completeness of their coding of diagnoses and procedures, which sometimes affects payment. It is difficult to distinguish these situations from genuine underlying changes in case mix.

SUMMARY AND CONCLUSION

Administrative data are usually by-products of health care delivery and financing systems, which means they are less costly to obtain, but they do not always coincide with research needs. Nonetheless, administrative data represent a relatively simple and readily available source of comprehensive cost data for large numbers of individuals, covering long periods of time. The widespread use of common administrative datasets also permits comparisons of economic analyses across studies.

There is no single best source of administrative data; the best source for a given study depends on the study's purposes and timing. Nonetheless, some conclusions can be drawn about using administrative data to estimate costs. Payment data are appropriate for analyses of expenditures under public programs such as Medicare and Medicaid, as well as expenditures under private insurance plans. Payments are also a measure of provider revenue for specific services. Cost to charge ratios can more accurately measure production costs because they incorporate actual costs as reported by providers. Billed charges are a less satisfactory measure of costs, but they can serve as a proxy if more accurate measures are not available.

Administrative data have become an essential source of information for studies of the financial burden of disease. Policymakers depend on such studies to inform decisions on health care financing and other matters. As administrative data systems improve, particularly through linkage to other datasets, they will increasingly provide vital support to the development of national health policy.

ACKNOWLEDGMENT

The author wishes to thank David Gibson for his helpful comments on an earlier draft of manuscript. The statements contained in this paper are those of the author and do not necessarily reflect the views or policies of the Centers for Medicare and Medicaid services.

REFERENCES

1. Catlin A, Cowan C, Hartman M, et al. National health spending in 2006: a year of change for prescription drugs. *Health Aff*. 2008;27:14–29.
2. Agency for Healthcare Research and Quality. Available at: http://www.hcup-us.ahrq.gov. Accessed July 9, 2008.
3. Friedman B, De La Mare J, Andrews R, et al. Practical options for estimating cost of hospital inpatient stays. *J Health Care Finance*. 2002;29:1–13.
4. Centers for Medicare and Medicaid Services. Health Care Financ Rev Stat Suppl. 2007. Available at: http://www.cms.hhs.gov/MedicareMedicaidStatSupp. Accessed May 8, 2008.
5. Schwartz M, Young DW, Siegrist R. The ratio of costs to charges: How good a basis for estimating costs? *Inquiry*. 1995–96;32:476–481.
6. Lave JR, Pashos CL, Anderson GF, et al. Costing medical care: using Medicare administrative data. *Med Care*. 1994;32:JS77–JS89.
7. Research Data Assistance Center. Using Medicare hospital cost reports cost-to-charge ratios in research. Technical note. Available at: http://www.resdac.umn.edu/Tools/tech_pubs.asp. Accessed March 28, 2008.
8. Barnett PG. Review of methods to determine VA health care costs. *Med Care* 1999;37:AS9–AS17.
9. Warren JL, Klabunde CN, Schrag D, et al. Overview of the SEER-Medicare data. *Med Care*. 2002;40:IV-3–IV-18.
10. United States Renal Data System. Available at: http://www.usrds.org. Accessed March 31, 2008.
11. Gold HT, Do HT. Evaluation of three algorithms to identify incident breast cancer in Medicare claims data. *Health Serv Res*. 2007;42:2056–2069.
12. Riley GF, Potosky AL, Lubitz JD, et al. Medicare payments from diagnosis to death for elderly cancer patients by stage at diagnosis. *Med Care*. 1995;33:828–841.
13. Brown ML, Riley GF, Potosky AL, et al. Obtaining long-term disease specific costs of care. *Med Care*. 1999;37:1249–1259.
14. Etzioni R, Riley GF, Ramsey SD, et al. Measuring costs: administrative claims data, clinical trials, and beyond. *Med Care*. 2002;40:III-63–III-72.
15. Federal Interagency Forum on Aging Related Statistics. *Older Americans 2008: Key Indicators of Well-Being*. Washington, DC: US Government Printing Office; 2008.
16. Riley G. Risk adjusting capitation payments to health plans that disproportionately enroll frail Medicare beneficiaries. *Health Care Finance Rev*. 2000;21:135–148.
17. Taylor DH, Hoenig H. Access to health care services for the disabled elderly. *Health Serv Res*. 2006;41:743–758.
18. Eppig FJ, Chulis GS. Matching MCBS and Medicare data: the best of both worlds. *Health Care Finance Rev*. 1997;18:211–229.
19. Centers for Medicare and Medicaid Services. Chronic Condition Warehouse. Available at: http://65.117.255.59. Accessed March 28, 2008.
20. Crystal S, Akincigil A, Bilder S, et al. Studying prescription drug use and outcomes with Medicaid claims data. *Med Care*. 2007;10:S58–S65.
21. Warren JL, Feuer E, Potosky AL, et al. Use of Medicare hospital and physician data to assess breast cancer incidence. *Med Care*. 1999;37:445–456.
22. Seiber EE. Physician code creep: evidence in Medicaid and state employee health insurance billing. *Health Care Finance Rev*. 2007;28:83–93.
23. Hsia DC, Krushat WM, Fagan AB, et al. Accuracy of diagnostic coding for Medicare patients under the prospective-payment system. *N Engl J Med*. 1988;318:352–355.