Use of Medicaid and Medicare Administrative Claims Data in Litigation and Regulation

Timothy Wyant and Stephen T. Parente[1]

Decipher, Inc., 17644 Raven Rocks Road, Bluemont, VA 20135, tw@deciph.com and
University of Minnesota, 321 19th Street South, Minneapolis, MN 55455, sparente@csom.umn.edu

Introduction

In 1994, Mississippi and Minnesota independently filed lawsuits against the tobacco industry seeking reimbursement of smoking-attributable Medicaid expenditures. Numerous other states filed similar suits over the next several years. During 1997-1998 the original pair, plus Florida and Texas, settled their suits during or on the eve of trial. In November 1998, the remaining 46 states negotiated the "Master Settlement Agreement" (MSA) with the tobacco industry. The MSA imposed various advertising and marketing prohibitions on the industry, and the industry agreed to pay $206 billion to these 46 states over the next 25 years.

In 1999, the United States Department of Justice filed a similar suit, seeking reimbursement of smoking-attributable Medicare expenditures under the Medicare Secondary Payor Act (MSPA). The MSPA suit was dismissed in 2001, although a related RICO suit is still ongoing. (The RICO suit does not involve reimbursement of Medicare expenditures.)

The Medicaid and Medicare suits required estimates of past smoking-attributable expenditures by these programs over periods of up to 20 years. In most of the suits, administrative claims data were used to calculate such quantities as:

- Total program expenditures by age, gender, and expenditure category (e.g. nursing homes v. hospitals)
- Total annual expenditures for medical encounters to treat specific smoking-related diseases, such as lung cancer, emphysema, coronary heart disease, or stroke
- Number of beneficiaries being treated for smoking-related diseases, by calendar year
- Total annual cost of treating beneficiaries with smoking-related disease, including not only medical encounters for direct treatment of the diseases, but also encounters for treatment of complications or for treatment of other conditions made more difficult or more expensive to treat (e.g. surgeries made more difficult or dangerous due to poor respiratory health)

Administrative claims data offer obvious advantages in these settings – they contain the most accurate and complete information on the outcome of interest (payments for treatment), and are directly relevant to the programs that are central to the lawsuits. They offer additional advantages in regulatory and surveillance settings in that the datasets are large and comprehensive, the data have been collected in a reasonably consistent manner over a number of years, and the data will continue to be collected, using similar procedures, into the future. On the downside, there are some challenges to working with claims data. We will describe some of them below, and describe some approaches that have proved useful.

We worked on statistical estimates of smoking-attributable expenditures in the Medicaid lawsuits on behalf of the Attorneys General for Minnesota, Maryland, and Wisconsin, and on behalf of the U.S. Department of Justice in the Medicare suit. Detailed descriptions of statistical modeling of smoking-attributable expenditures can be found in a number of places, for

[1] We thank Jonathan M. Samet, M.D., M.S. for his suggestions regarding clinical significance of ICD-9 codes and sequelae of smoking-attributable conditions.

example Coller (2002) and Max (2001). Here, we focus on some specific issues related to the use of administrative claims data in such analyses.

Nature of the data

Administrative claims data for health care programs are the electronic versions of bills submitted by physicians, hospitals, pharmacies, or other medical providers for office visits, hospital stays, or other encounters, or for sales of drugs or supplies. There are reasonably standard billing forms and procedures that most providers and payors use for billing and payment. The word "reasonably" suggests (as is so often the case) numerous idiosyncrasies and anomalies in billing systems that, while not of great importance individually, can generate annoyance and expense for the analyst. These departures from a simple uniform standard are surmountable with sufficient (albeit at times nontrivial) efforts by programmers and analysts, and we will not try to give any kind of comprehensive summary in this article.

In most large health care programs, these electronic claims records are actually used to cut and send checks to the providers. Consequently, there is a profound ongoing incentive for many parties (the program, oversight groups and auditors, the providers, and the beneficiaries) to make sure that the basic information in these data is accurate and up-to-date. This inherent quality control is an appealing aspect of these data; the quality control for other kinds of administrative data that are maintained for "informational" purposes tends to wax and wane with changes in budgets and leadership.

Records in the administrative claims data contain information on items such as 1) date and location of service, 2) type and cost of service, 3) procedures performed, 4) extent of service (e.g. hospital days), 5) beneficiary demographics such as age, gender, and location of residence, and 6) program information for the beneficiary, such as type and dates of coverage, and information needed for billing and mailing purposes (addresses, phone numbers).

To preserve confidentiality, the latter information – addresses and the like – is typically purged from or encrypted in administrative claims data that are extracted for analytic purposes. However, sufficient information is usually maintained to allow the analyst to group the encounters by beneficiary.

In addition, and of particular importance for attributable cost studies, encounter records contain the diagnoses associated with the physician visit or hospital stay. These are coded using the International Classification of Diseases, Version 9 (ICD-9). [2]

Administrative claims data for Medicare reside in the Center for Medicare and Medicaid Services' (CMS) National Claims History Files (NCHF). Administrative claims data for Medicaid programs typically reside in the Medicaid Statistical Information System (MSIS), although there are states that do not participate in the MSIS data system.

One issue that typically arises in these administrative claims data is that services provided through managed care contracts are often not covered. For such services, Medicare and Medicaid programs pay lump sum fees to the managed care contractors, who often do not keep cost records on encounters in the same way that "fee-for-service" providers do. However, the fee-for-service records that do reside in the administrative claims data typically make up the bulk of claims activity in the Medicare and Medicaid programs, and there are usually straightforward ways to extrapolate from these data to the managed care programs if it is necessary to do so.

As one example of the scope of claims data analyzed for the tobacco litigations, we processed 224 million claims records in estimating smoking-attributable health care expenditures by the Minnesota Medicaid program over a 15-year period.

[2] ICD codes were originally developed to code cause of death, but have become the standard for billing in the U.S. The World Health Organization (WHO), as of this writing, has moved on to ICD-10 codes as the standard for research purposes, but the medical billing world still runs on ICD-9. WHO transferred the license for use of ICD-9 to CMS, so that this version of the classification system can continue to be used for billing and reimbursement purposes under the aegis of a recognized authoritative body.

The core calculation of smoking-attributable costs

For simplicity, look at one disease, in one year, for one gender and one age. For example, we can look at COPD (chronic obstructive pulmonary disease) for 50 year-old males in 2001. Assume there are *N* beneficiaries who are 50 year-old males in 2001, indexed by *i = 1, 2, …, N.*

COPD can be defined as any of the following diagnoses, expressed as ICD-9 codes:

Chronic obstructive pulmonary disease (COPD)	491 Chronic bronchitis 492 Emphysema 496 Chronic airway obstruction, not elsewhere classified

The core calculation is:

1) Calculate the total health care costs C_i during the year for each beneficiary in the group of 50 year-old males. "Costs" in this context are the total payments made by the Medicaid or Medicare program to reimburse medical providers for treating the beneficiary during 2001.

2) Identify those beneficiaries in the group that are treated for COPD sometime during the year. Assume that subgroup *D* contains the N_D beneficiaries who were treated for COPD.

3) Obtain the probability *P(S)* that a 50 year-old male will be a smoker. This probability is typically obtained from surveys of either beneficiaries of the health care program, or people similar to those beneficiaries.

4) Obtain the probability that a 50 year-old male smoker will be treated for COPD during a calendar year. Call this probability *P(D|S)*. Typically, *P(D|S)* will come from some external study that quantifies the extent to which smokers have an elevated risk of being treated for COPD.[3] In a similar fashion, obtain *P(D|~S)*, where *P(D|~S)* is the probability that a non-smoker will be treated for COPD during a calendar year.

5) Calculate the probability that a 50 year-old male with COPD is a smoker:

$$P(S \mid D) = \frac{P(D \mid S)P(S)}{P(D)} \quad \text{where} \quad P(D) = \frac{N_D}{N_D + N_{\sim D}}$$

6) Calculate the average expenditures per beneficiary during the year for 50 year-old males with and without COPD:

$$C(D) = \frac{\sum_{i \subset D} C_i}{N_D} \qquad C(\sim D) = \frac{\sum_{i \not\subset D} C_i}{N_{\sim D}}$$

7) Calculate the total smoking-attributable expenditures for COPD in this group as:

$$SAE = \sum_{i \subset D} P(S \mid D) \ \frac{P(D \mid S) - P(D \mid \sim S)}{P(D \mid S)} \ \frac{C(D) - C(\sim D)}{C(D)} \ C_i$$

[3] In some instances, if the health care program conducts a survey of its members to determine which ones smoke, *P(S)* and *P(S|D)* can be calculated using administrative claims data. The Medicare Current Beneficiary Survey (MCBS) can be used for this purpose when analyzing Medicare data.

In other words, we sum over all the 50 year-old males who are being treated for COPD. For each such beneficiary, we calculate the total costs C_i. However, C_i is not smoking-attributable in its entirety, because the hypothetical beneficiary a) might not be a smoker, b) might have gotten COPD even if he hadn't smoked, and c) would likely have incurred some costs even if he hadn't been treated for COPD. The terms to the right of the summation sign in the above SAE formula adjust for each of these possibilities in turn.

The smoking-attributable fraction of expenditures can be defined as:

$$SAFE = \frac{SAE}{\sum_{i=1,N} C_i}$$

An overall SAE for the health care program can be calculated by adding up the SAEs for each age, gender, calendar year, and disease.[4] An overall SAFE can be calculated by dividing the overall SAE by the total program expenditures for the years in question. [5]

Special concerns when using administrative claims data to quantify attributable costs

Identifying diseases in claims data

Disease-specific cost studies using administrative claims data must identify the encounters that "map to" the disease in question. This is typically done by first defining the disease with one or more ICD-9 codes (as in the COPD example above), and then passing through the encounters in the claims data, performing a look-up of each encounter's diagnoses.

The potential pitfall in this approach is that disease definitions obtained from the medical literature arise out of "cause of death" ICD-9 codings. Codings for billing and reimbursement may differ.

In examining smoking-attributable disease, for example, the Surgeon General would typically define "stroke" or "cerebrovascular disease" as "codes in the 430s":

Cerebrovascular disease	430 Subarachnoid hemorrhage
	…
	438 Late effects of cerebrovascular disease

Examination of computer-generated summaries of claims histories for individuals suggests that, in many health care programs, the most common diagnosis coding for stroke is "one-sided paralysis:"

342 Hemiplegia and hemiparesis

[4] Here, for simplicity, we ignore issues related to how one classifies a beneficiary who has more than one smoking-attributable disease. There are a number of effective approaches, including imposition of a relative risk-based hierarchy of disease, or multivariate modeling of cost as a function of disease(s).

[5] Because the focus of this article is on analysis issues related to the use of administrative billing records, we have posited a simple formulation of smoking (yes/no), and of models relating costs to disease, and disease to smoking. Many analysts have explored more complicated approaches, expressing smoking as some function of current/former, age of initiation, duration of smoking, and total "dose" in pack-years, and using various functional forms and multivariate models to model the other relationships. Nonetheless, the basic ideas are expressed in the core calculation presented above. And in fact, the relationship between smoking and diseases such as lung cancer, and between diseases such as lung cancer and costs, are so strong and direct that simple formulations are often quite adequate.

This makes sense from a treatment and billing perspective – the stroke itself might have occurred years ago, there is nothing that can be done to reverse it. What the provider is treating today is problems related to the paralysis – the ongoing result of the original stroke that is in turn causing other health problems. Relying on the Surgeon General-type coding would miss many instances of stroke-related treatment.

Identifying people being treated for a disease: avoiding false positives

Diagnoses recorded in administrative claims data encounters are not perfect. Because of mistaken or missed diagnoses, or data entry errors, there are both false positive and false negative errors of disease identification.

There are two reasons an analyst, in examining administrative claims records, might be more concerned with the "false positive" disease identifications. One reason might be termed strategic; the other might be termed structural.

The strategic reason is that in some instances there is an asymmetric loss function – one would prefer to err on the side of understating, rather than overstating, costs associated with a particular disease or health behavior. This was typically the case for analysts who worked for plaintiffs in the smoking-attributable cost litigations. The smoking-attributable costs were generally so huge that there was no point in wasting the court's time dealing with disputes over the best way to estimate small marginal contributions to attributable costs.

The structural reason relates to an artifact of the medical billing system in the U.S. There is no way on the standard billing forms to indicate that a procedure was intended to "rule out" or "check for" a possible disease, as opposed to treating a confirmed disease. So, for example, if a doctor performed a bronchoscopy on a heavy smoker with chest pain to check for lung cancer, the billing form would likely have an ICD-9 code for "lung cancer" even if the test proved to be negative. Without the lung cancer designation, the health plan might refuse to reimburse the procedure on the basis that it was "unnecessary".[6]

A number of strategies have been employed, singly or in combination, to reduce the chance of false positives. These include 1) restricting disease identifications to primary (as opposed to secondary) diagnoses, 2) restricting identifications to hospital encounters, 3) ignoring encounters in which only diagnostic procedures were performed (e.g. radiology, laboratory, pathology), 4) requiring multiple diagnoses of the disease at different encounters separated by some amount of time – Quam (1993), and 5) imposing some credibility restrictions – for example, in our smoking-attributable cost studies, we have typically ignored diagnoses of stroke and lung cancer for beneficiaries under age 40. Such conditions occur so rarely at those ages that there is an elevated chance that the diagnosis in the claims data is mistaken.

These strategies can be used in various combinations. For example, in smoking-attributable cost studies, we have identified treatment for a disease in a year if 1) there is a primary diagnosis of the disease associated with a inpatient hospital stay or 2) there are diagnoses on at least two physician visits in different quarters, where the visits were not exclusively for diagnostic testing.

Using total health costs, rather than costs for selected encounters

One approach to using administrative claims data in smoking-attributable cost studies is to first identify only those encounters that have diagnoses of smoking-attributable diseases. The costs for these encounters are then taken as the maximum smoking-attributable costs. These costs are then further reduced according to the probability that 1) the beneficiary is a smoker, and 2) if the person is a smoker, that the encounter is due to the smoking. All encounters without a diagnosis for the diseases in question are ignored.

Although this approach has some intuitive appeal, and can be useful in some circumstances to calculate a lower bound on attributable costs, it generally is too flawed to be of practical use.

[6] In smoking-attributable cost studies, elevated testing rates for smokers would in fact be a legitimate component of smoking-attributable costs, but because of the desire for simple and conservative estimates that focus on specific diseases, this "excess testing" component is often ignored.

This is because this "disease-specific encounter" approach ignores all the health consequences – and associated costs – of having diseases such as cancer, COPD, or coronary heart disease. In administrative claims data, many of these consequences will show up in encounters that do not have the diagnoses of the original diseases. Simple examples are metastasizing lung cancer, or surgical complications. If a person gets brain cancer that is secondary to an original lung cancer, then encounters to treat the brain cancer will have brain cancer diagnoses in the claims data. If a person gets an infection in a surgery for lung cancer, subsequent hospitalizations to treat the infection will have infection diagnoses in the claims data. If the analyst looks only at encounters that have a "lung cancer" diagnosis, he or she will miss costs for subsequent metastases or infections that are clearly a consequence of the lung cancer.

But such examples are only the tip of the iceberg. Having a background condition such as COPD makes other diseases and conditions more difficult or costly to treat. Surgeries may be postponed because of elevated risk, more expensive drugs may be prescribed to avoid harmful interactions with drugs prescribed to treat the COPD, respiratory infections may require hospitalization instead of being treated at home. In addition, there are oddities of the billing and reimbursement system in the U.S. to take into account. If a person goes in for a series of chemotherapy treatments for lung cancer, the treatments are likely to appear in claims data not with a lung cancer diagnosis code, but with a "V" code indicating ongoing treatment for a previously diagnosed condition. Focusing solely on encounters with a lung cancer diagnosis would miss such clearly related "V" code encounters.

Our approach is to first identify beneficiaries who are being treated sometime during the year for diseases such as lung cancer, then look at *all* their costs for the year. We adjust for costs that would have occurred anyway by looking at costs for a "control group" of beneficiaries of similar age and gender who are not treated for lung cancer, and subtract them out. This approach captures in a reasonable way *all* of the costs associated with having a disease such as cancer, COPD, or coronary heart disease, and is consistent with the "control group" approach of other cost of disease studies. For example, see Fireman (1997).

Results and future work

The CDC estimates that about 19% of annual deaths in the U.S. are attributable to smoking (McGinnis (1993)). Typical estimated percentages of U.S. health care costs attributable to smoking fall in the 6-14% range (Max (2002)). However, Medicaid and Medicare programs can have smoking-attributable cost percentages that differ from the national percentages because they serve populations that are sicker and more prone to smoke (Medicaid), or are elderly or disabled (Medicare). Administrative claims data from these programs can provide a reliable basis for estimating smoking-attributable costs specific to the programs. For example, our estimates for Minnesota Medicaid were 11% for medical services and 3% for nursing home fees. This case settled for $6.1 billion during final arguments to the jury.

We are currently using Medicare administrative claims data to quantify rates of increased health care cost and utilization in response to elevated short- or long-term air pollution levels.[7]

References

Coller, Maribeth (2002), with Glenn Harrison, and Melayne McInnes, "Evaluating the Tobacco Settlement Damage Awards: Too Much or Not Enough?," *American Journal of Public Health*, 92(6): 984-989.

Fireman, Bruce H. (1997), with Charles P. Quesenberry, Carol P. Somkin, Alice S. Jacobson, David Baer, Dee West, Arnold L. Potosky, and Martin L. Brown, "Cost of Care for Cancer in a Health Maintenance Organization", *Health Care Financing Review*, 18(4): 51-76.

Max, Wendy (2001) "The financial impact of smoking on health-related costs: A review of the literature," *J Health Promotion*, 15(5): 321-331

[7] EPA Star Grant RD83054801.

McGinnis, JM (1993), with W.H. Foege, “Actual causes of death in the United States”, Journal of the American Medical Association, 270(18):2207-2212.

Quam, L. (1993), with Ellis, L. B.; Venus, P.; Clouse, J.; Taylor, C. G.; Leatherman, S., “Using claims data for epidemiologic research. The concordance of claims- based criteria with the medical record and patient survey for identifying a hypertensive population”, *Med Care,* 31(6):498-507.