Published in final edited form as:
Am J Med Qual. 2006 ; 21(4): 269–275.

Challenges of Using Medical Insurance Claims Data for Utilization Analysis

Patrick T. Tyree, AA[1], **Bonnie K. Lind, PhD**[1,2], and **William E. Lafferty, MD**[1]

1 *Department of Health Services, School of Public Health and Community Medicine, University of Washington*

2*Department of Nursing, Boise State University*

Abstract

Research use of insurance claims data presents unique challenges and requires a series of value judgments which are intended to improve the data quality. In our sturdy, medical insurance claims from two large companies were combined to assess utilization of complementary and alternative medicine. Challenges included assessing and improving the quality of data, combining data from two different companies with dissimilar coding systems, and determining the most appropriate ways to describe utilization. This paper addresses four methodologic challenges in creating the analytic files: (1) conversion of claims into unique visits, (2) identification of incomplete claims data, (3) categorization of providers and locations of service, and (4) selecting the most useful measures of utilization and expenditures.

Keywords

Administrative data; insurance claims; utilization; expenditures; methods; provider type; location of service; data quality

INTRODUCTION

Health insurance administrative data are an important source of information for medical research. Analyses of insurance claims data were reported in at least 200 published articles in 2004. They are a rich and relatively inexpensive source of research information for studies of health care utilization and medical expenditures. Claims data record diagnostic information, treatments given, and providers used, in addition to a variable number of financial measures such as billed amounts, reimbursed amounts, and patient cost sharing. Claims data have been shown to exhibit high congruence with medical records data compared to patient surveys, both telephone and mail.(1, 2) However, because these data were developed for administrative purposes the conversion of claims into a research database requires substantial effort.(3) We faced this challenge in the Insurance Financing of Integrative Medicine (IFIM) study, (4–7) funded by the National Institutes of Health's National Center for Complementary and Alternative Medicine. This study used medical claims data from two large insurance companies to assess the utilization of complementary and alternative medicine (CAM) over three different twelve month time periods (1997, 2000, and 2002). Converting claims data into analytic files

Corresponding author : William E. Lafferty, MD, Department of Health Services, School of Public Heath and Community Medicine, University of Washington, 1959 NE Pacific St, Box 357660, Seattle, WA 98195-7660; billlaf@u.washington.edu.

This publication was made possible by Grant Number NIH R01 AT000891-05 from the National Center for Complementary and Alternative Medicine (NCCAM). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Complementary and Alternative Medicine, National Institutes of Health.

presented challenges related to assessing and improving the quality of the data, combining data from two different companies, and determining the most appropriate ways to describe utilization. This paper focuses on four methodologic challenges in creating the analytic files: (1) the creation of a unique visit for analyzing utilization rates by provider type, (2) identification of incomplete claims data, (3) categorization of providers and locations of service, and (4) selecting the most useful measures of utilization and expenditures.

METHODS

The study was approved by the University of Washington's Institutional Review Board. Confidentiality agreements were established between each of the two insurance companies and the University of Washington prior to data sharing. Three years of medical insurance enrollment and claims data were requested from each insurance company based on the study eligibility criteria (Washington State residents, age birth through 64, who were not covered by publicly funded Medicare or Medicaid, and were not Federal employees). All records were assigned an anonymous enrollee specific identifier by the insurance company.

Four relational files came from each company. The enrollment file had one line of data for each unique enrollee and included data such as age, gender, and zip code of residence. The claims file, which linked to the enrollment file by a patient identifier, had one line of data for each procedure done at each visit. A provider file, which linked by provider identifier to the claims data, supplied practitioner type and billing zip code. The pharmacy file had one record for each enrollee and linked to the enrollment file by patient identifier. It supplied aggregate annual pharmacy expenditure and the annual number of prescriptions filled.

Each participating company assigned a medical director to be the contact for the study. The study investigators developed a comprehensive group of dummy tables that specified variables, methods of aggregation, and permissible non-identifying cell sizes which were approved by the medical directors prior to starting analysis. Neither of the companies required final approval of manuscripts or specific findings as a condition for participation.

During the IFIM study we kept a log-book of all the database manipulations and what effect they had on study results. Because we did not review medical records, we used a series of indirect analyses to assess completeness and quality of these data. In addition, enrollment files were checked against specified eligibility criteria. Enrollees not meeting eligibility criteria were excluded from the final analysis database. Both companies' enrollment and claims file variables were checked against their respective data dictionaries for values that were missing or outside of defined ranges. The insurance companies were consulted to resolve discrepancies. In most cases the data dictionary was updated to reflect the additional acceptable values. In a few cases, the data values represented a temporary coding scheme and were updated to reflect a database dictionary defined permanent value.

RESULTS

Data Quality Assessment

Potential threats to data quality occur in the process of converting medical chart data to claims and during subsequent claims processing.(8) As a proxy for claim billing errors, gender specific conditions in the adult claims data were compared to the gender designation in the enrollment files. Male conditions included diseases of the male genital organs (ICD-9(9) 600–608); female conditions included inflammatory disease of female pelvic organs (614–616), other disorders of female genital tract (617–629), and complication of pregnancy, childbirth, and the puerperium (630–676). Discrepancies between gender and the presence of these conditions over the three years averaged only 0.27% with no notable rate difference between the two companies. Given the focus of our project, a separate analysis restricted to outpatient visits compared these rates for CAM providers versus conventional providers. During the three year period outpatient claims from conventional providers averaged a 0.28% discrepancy rate with a 0.45% rate from CAM providers.

Creation of a Unique Visit

We chose to evaluate provider utilization by a metric of the number of unique visits. Creating unique visits required two procedures, since claims data are organized around a billing event rather than an actual visit. First, several visits may be submitted on one claim, particularly from providers that provide repetitive services, e.g. a chiropractor, massage therapist, or a psychologist. Thus, the unique identifier on each claim does not necessarily identify a unique visit. These visits must be separated because if the claim number were allowed to represent a unique visit, the actual number of visits would be undercounted. For one insurance company approximately 6% of their outpatient claim numbers represented batched visits (3% of their total claims). This ranged from 1.6% of claims for conventional providers to 26% of claims for CAM providers. The rate was substantially lower for the other company. Understanding this relationship between claims and visits and how they relate to billing is important. Historically we have been able to use the unique claim number to establish this relationship. However, under the Health Information Portability and Accountability Act (10) claims numbers are considered direct identifiers and as such, may no longer be available to researchers. In this case, date of services and the provider identifier could be used to establish unique visits without using a claim identifier.

The second process involved identifying and combining multiple claims from a single visit. The claims data contain one line for each billed procedure; an individual visit is represented by a grouping of billed procedure codes. A visit was defined as one encounter to any given provider per day. Each line of data was assigned a visit identifier based on a unique patient identifier, provider identifier, and date of service. In the final analysis dataset, each procedure code (CPT(11) and HCPCS(12)) was kept only once per visit regardless of the number of times it appeared in the data. Although a procedure may have been performed multiple times, billing protocols dictate that a procedure is only listed once. The number of times a procedure is performed is maintained by the insurance companies in a variable named “units”.(13, 14) For example, a 45 minute therapeutic massage should be coded as CPT 97124 (therapeutic procedure, 15 minutes, therapeutic massage) with a units designation of 3. Therefore, if there were multiple records with the same CPT or HCPCS codes for a single visit, the duplicates represented an error and were deleted.

Billing adjustments accounted for the majority of duplicate records. Adjustments were made in two ways: (1) a bill being submitted twice with different claim reference numbers, and more commonly (2) claims data containing both the original claims and adjustments to those claims. For example, a line of data might show an original request for payment followed by a second line containing a negative monetary amount, thus wiping out the expenditures listed in the first claim; the third line would be the reentry of the claim. This often occurred with different claim numbers assigned.

Duplicate data also might conceivably occur due to a second visit to the same provider on the same day. Without chart review it is impossible to tell whether multiple lines of duplicate procedure data represent a second visit on the same day or an adjustment to the original data. However, multiple visits to a single provider on a given day should be rare events, so these records were always treated as adjustments.

In cases where procedures were listed more than once, the highest unit and monetary values were maintained. This included the individual maximum amounts listed on any line for billed, allowed, paid, co-pay, co-insurance, deductible, and coordination of benefits. The duplicate billings were then eliminated. In some cases it appeared that the enrollee had a second visit during the day based on different procedure codes. In such a case all unique procedures were attributed to a single visit. The exceptions to this rule are the CPT procedure codes for

Evaluation and Management (99201–99215) which designate length of visit and whether the visit was for a new or established patient. If more than one of these procedures are present, the code representing the highest level of service was maintained.

Detecting Incomplete Claims Data by a Contract Level Analysis

Overall utilization rates were calculated as the proportion of enrollees with at least one inpatient or outpatient claim during the year. On initial analysis, the annual utilization rates ranged from 67.3% in 1997 to 71.0% in 2002. Because these utilization rates were lower than previous reports,(15, 16) we did further exploration of the data to see if any data anomalies were present to explain this. We found a significant number of contracts for which fewer than 50% of enrollees had any claim during the year. (The 50% figure was arbitrary, but based on the previously reported utilization rates cited above, it was a conservative cutoff for the range.)

Further discussion with the insurance companies revealed that these contracts, which we labeled "incomplete contracts," arose primarily from insurance company mergers in which pre-merger claims were maintained separately and not available for our analysis. A second source of incomplete contracts was claims being contracted out for management by an independent company. Thus, we deemed that these incomplete contracts in fact represented incomplete data and should be excluded from further analysis. However, in contracts with only a few enrollees, the utilization rate might be expected to be low, by chance. Therefore we only considered a contact as incomplete if utilization was below 50% and the contract had at least 10 enrollees. Contracts with fewer than 10 enrollees accounted for about 10% of all enrollees each year and were considered complete regardless of utilization. After excluding incomplete contracts, the overall range of utilization increased to 78% – 83%. Table 1 displays the percentages of contracts and of enrollees affected by incomplete contracts.

Developing a Strategy to Unify Variable Categorization

The coding of demographic variables (e.g., date of birth, gender, and zip code) showed little variation between the two insurance companies. Claims data variables, such as provider type and service location, required standardization prior to analysis of a combined database. This was particularly important for grouping provider types, where each company had over 60 provider categories and twice as many specialty codes, many of which were defined differently by the two companies. The two insurance companies also defined location of service categories somewhat differently, so we intentionally created broad major categories such that all current and future values could be accommodated in a consistent categorization. As shown in Table 2, general provider types could be grouped into five major categories. Table 3 shows the categorization of location of service. The original provider and location designations were maintained, for reference.

Further Clarification of Provider Type

Both insurance companies had several providers listed under the provider type categories of Clinic or Miscellaneous. Efforts were made to classify these into their actual provider types. Each provider name was reviewed for an appropriate classification. Clinics which included dental, eye, optometry, or pharmacy in their names would have their associated claims excluded from the analysis file. This method was equally important in classifying CAM providers who had naturopathic, acupuncture, massage, or chiropractic clinic in their names. Where exclusion or categorizing to CAM provider types was inappropriate, the provider type was classified as *all other*.

Determinants of Insurance Utilization and Expenditures

Rates of utilization based on the mere presence of a claim would not represent actual expenditures (enrollees and providers may submit claims for non covered services and duplicate billing may inadvertently occur). Therefore, our study reviewed associated pecuniary data to make determinations of utilization for covered benefits. Only visits that were allowed by the insurance companies (allowed amount greater than zero) were considered as a utilized service. Allowed amounts were chosen because the objective of the IFIM project was to look at the effect of insurance coverage of CAM providers on utilization and expenditures under insurance. Thus for this project we were only interested in claims which affected the insurance company; other projects which use claims may define utilization differently based on the objectives of those projects. There are few claims omitted with the allowed amounts method; in 2002, 96% of all visits to an outpatient conventional provider were allowed.

The allowed amount was not only used as the marker for insurance utilization, but it was also chosen to represent the expenditures for each visit. In addition to its unique characteristic of representing the covered benefit, it had several advantages over using billed or paid amounts. Billed amounts were highly variable between providers of a given type and did not necessarily reflect the amount the insurance company was contracted to cover. Further, the billed amount may or may not be for a service covered by the benefits contract. Allowed amounts demonstrate less variability; as shown in Table 4, the standard deviation from the mean is lower in allowed versus billed amounts. Paid amounts only reflect the amount the insurance company pays and are dependent on patient cost sharing requirements such as deductibles, co-payments, and co-insurance amounts. Allowed amounts are the best indicator of the overall cost of the services to the patient and the insurance company. In virtually all cases, the difference between the amount paid and the amount allowed was attributable to deductibles, co-payments, and coinsurance (the allowed amount being higher because it includes the patient cost-sharing amounts). When expenditures for 5 common procedures were compared, conventional providers received 63% of mean billed to allowed, whereas CAM providers received 80%.

The use of allowed amounts required adherence to strict rules about the presentation of these data in order to maintain confidentiality of company specific contracted rates. As described above, the expenditure data from all individual services were combined into visit level data. These data were presented only when aggregation across large numbers of services and patients assured that the original figures could not be derived. Thus, sensitive proprietary information on specific negotiated rates of payments to individual providers, specific provider groups, and the expenditures for specific services were never reported.

CONCLUSIONS

Research use of insurance claims presents unique challenges and requires a series of value judgments which are intended to improve the data quality. Other judgments could have been made, thus establishing different results. To ensure reproducibility, it is important for researchers to include, in their papers or appendices, how the particular issues raised in our methods section were handled.

Insurance companies must have a high degree of trust in their research partners to allow use of proprietary data because of the remote chance of inappropriate release. Our study was fortunate to have partners that were supportive of the need for health services research and who understood the risks and benefits of this work. Under different circumstances, an insurance company may choose to protect their proprietary information by not releasing billed and allowed data. In the absence of these expenditures, linking CPT, HCPCS, and diagnosis related groups (DRGs) to standard rates of reimbursement, such as the Medicare fee schedule, would be a recommended alternative approach.

HIPAA has heightened concerns over data use, however most research may still proceed with appropriate safeguards. For example, the individual claims identifier is now considered confidential data. Our method of aggregating and splitting all claims by visit date and provider obviates the need for this information.

We found that aggregating data by contracts and performing completeness checks was essential. Insurance companies are subject to frequent mergers and acquisitions, thus we found the opportunity for incomplete data is substantial and must be investigated. Presenting the companies with a list of potentially incomplete contracts validated our suspicion that most of these represented missing utilization data instead of bona fide instances of low insurance use.

Our data requests specifically restricted the population of interest to enrollees with 12 months of continuous coverage, thus we expected complete data. Although complete data is desirable, there is no single source for this information at the insurance companies. Due to the disconnect between enrollment and claims information storage at the insurance companies, the enrollment data may be present in the file but the claims data may not have been assimilated in. Using our method of determining incomplete contracts is one way to limit the data to only enrollees with complete claims history.

Including alternative care providers into insurance reimbursement presented several challenges for data analysis. In our study, CAM providers tended to submit bills in batches and thus claims often represented multiple dates of service. We also found that provider categories were defined differently between companies. Thus, study-specific algorithms for combining like providers will always be necessary. The coverage of CAM services is relatively new, and we suspect that there are many CAM providers who don't accept insurance. Therefore, utilization rates derived from claims data are believed to be a minimum estimate of utilization.

Using claims data to investigate research questions is subject to several limitations. Claims data are dependent on professional ICD coding. In the clinical setting, some diagnoses may be missed, different professional types may have different coding patterns, and not all coding may be accurate. When using a multi-company combined claims database, variations in benefit structures between insurance companies may affect utilization analysis. Finally, it is important to recognize that utilization results gained from claims analysis apply only to the insured population.

Another limitation of claims data is the inability to assess outcomes. Clinical providers of CAM services who were members of our project's community advisory group were often interested in clinical outcomes. For a variety of reasons we cannot address outcomes questions using claims data. First, outcomes are not explicitly included in claims data, so outcomes information must be inferred from existing information. However, claims information lack data on severity and duration of illness prior to the diagnosed event. This limits the ability to compare patients with like illness. In addition, while we believe that information on claims such as ICD codes are generally accurate, they function best when evaluating large numbers of clients with similar conditions rather than in evaluating individual outcomes. Finally, our ability to do substantial error checks was very limited. The fact that gender was usually coded correctly was reassuring but hardly represents the level of assurance that would be necessary for a clinical outcomes study.

In spite of these limitations, claims data are a valuable resource for exploratory analyses of a variety of health services research questions, and their continued use for this purpose should be encouraged. We caution researchers, however, to be aware of the methodologic issues involved in using these data, in order that valid and reproducible results can be obtained.

References

1. Fowles JB, Fowler E, Craft C, et al. Comparing claims data and self-reported data with the medical record for Pap smear rates. Eval Health Prof 1997;20(3):324–42. [PubMed: 10183327]
2. Fowles JB, Fowler EJ, Craft C. Validation of claims diagnoses and self-reported conditions compared with medical records for selected chronic diseases. J Ambul Care Manage 1998;21(1):24–34. [PubMed: 10181337]
3. Motheral B, Brooks J, Clark MA, et al. A checklist for retrospective database studies--report of the ISPOR Task Force on Retrospective Databases. Value Health 2003;6(2):90–7. [PubMed: 12641858]
4. Lafferty WE, Bellas A, Corage Baden A, et al. The use of complementary and alternative medical providers by insured cancer patients in Washington State. Cancer 2004;100(7):1522–30. [PubMed: 15042688]
5. Bellas A, Lafferty WE, Lind B, et al. Frequency, predictors, and expenditures for pediatric insurance claims for complementary and alternative medical professionals in Washington State. Arch Pediatr Adolesc Med 2005;159(4):367–72. [PubMed: 15809392]
6. Watts CA, Lafferty WE, Baden AC. The effect of mandating complementary and alternative medicine services on insurance benefits in Washington State. J Altern Complement Med 2004;10(6):1001–8. [PubMed: 15673994]
7. Lind BK, Lafferty WE, Tyree PT, et al. The role of alternative medical providers for the outpatient treatment of insured patients with back pain. Spine 2005;30(12):1454–9. [PubMed: 15959379]
8. Studney DR, Hakstian AR. Effect of a computerized ambulatory medical record system on the validity of claims data. Med Care 1983;21(4):463–7. [PubMed: 6843199]
9. U.S. Department of Health and Human Services. International Classification of Diseases, 9th Revision, Clinical Modification: ICD-9-CM. Washington, DC: U.S. Government Printing Office; 2001.
10. Office for Civil Rights. OCR Privacy Brief: Summary of the HIPAA privacy rule. U.S. Department of Health and Human Services, Washington D.C. Available at http://www.hhs.gov/ocr/privacysummary.pdf Accessed on May 26, 2005.
11. American Medical Association. Current Procedural Terminology. Chicago, IL: American Medical Association; 2000.
12. American Medical Association. Healthcare Common Procedure Coding System. Chicago: American Medical Association; 2000.
13. Centers for Medicare and Medicaid Services. Health Insurance Claim Form, Form CMS-1500. U.S. Department of Health and Human Services, Washington D.C. Rev.12/1990. Available at http://new.cms.hhs.gov/cmsforms/downloads/CMS1500.pdf Accessed January 02, 2006.
14. Centers for Medicare and Medicaid Services. Medicare Claims Processing Manual: Chapter 26 – Completing and Processing Form CMS-1500 Data Set. U.S. Department of Health and Human Services, Washington D.C. Rev.10/2005. Available at http://new.cms.hhs.gov/manuals/downloads/clm104c26.pdf Accessed on January 2, 2006.
15. Pleis JR, Coles R. Summary health statistics for U.S. adults: National Health Interview Survey, 1998. National Center for Health Statistics. Vital Health Stat 2002;10(209)Available at: www.cdc.gov/nchs/data/series/sr_10/sr10_209.pdf. Accessed January 2, 2006.
16. Pleis JR, Benson V, Schiller JS. Summary health statistics for U.S. adults: National Health Interview Survey, 2000. National Center for Health Statistics. Vital Health Stat 2003;10(215)Available at: http://www.cdc.gov/nchs/data/series/sr_10/sr10_215.pdf. Accessed January 2, 2006.

Table 1

Incomplete Contracts

	Total		Percent with < 50% Utilization [a]		Utilization	
	Contracts	Enrollees	Contracts	Enrollees	Before excluding incomplete contracts	After excluding incomplete contracts
1997	19,669	818,099	6.1%	16.7%	67.3%	78.0%
2000	35,307	1,030,867	4.5%	15.4%	68.0%	79.2%
2002	37,841	932,756	2.2%	15.3%	71.0%	83.0%

[a] Restricted to contracts with at least 10 enrollees

Table 2

Provider Groups

Conventional:	Physician Advanced Registered Nurse Practitioner Physician's Assistant Doctor of Osteopathic Medicine *Conventional care providers can be further subcategorized into to those providing primary care vs. those providing specialty care* Primary Care providers include: Internal medicine, Family medicine, General practice, Pediatrics, Adolescent, Geriatrics, Preventive medicine, Osteopath, Women's health specialist (when held by an ARNP) Specialty Care providers include all specialties not detailed under Primary Care
CAM:	Acupuncturist Chiropractor Licensed massage therapist Naturopathic physician
Physical Therapist:	Physical therapist
All Other:	Everything not specified as Conventional, CAM, or Physical Therapists, including: Alcohol treatment center, Smoking cessation program, Ambulance, Audiologist, Blood bank, Biofeedback therapist, Certified Hypnotherapist, etc.
Excluded from Analysis File:	Pharmacists, Durable goods suppliers, Prosthetic suppliers, Dentists, Dental specialists, Denturists, Optometrists, and Opticians

Table 3

Location of Service

Inpatient:	
	Inpatient hospital
	Nursing home/Skilled Nursing Facility (covered under health insurance; is generally post-surgical care)
	Inpatient mental health
Outpatient Clinic and Provider Office:	
	Outpatient hospital clinic
	Provider office
Outpatient Other: All locations not specified above, including:	
	Emergency room, Hospice, Kidney dialysis center, Outpatient surgical ambulatory, Drug and alcohol treatment (inpatient or outpatient), Outpatient mental health, Patient's home, Laboratory, X-ray, Blood bank, Ambulance

Table 4

Difference for Average Billed and Allowed Amounts

	Billed			Allowed		
Procedure[a]	**Mean**	**S.D.**	**Median**	**Mean**	**S.D.**	**Median**
1	90.18	41.46	86.80	54.18	23.03	43.01
2	36.18	7.09	35.00	27.22	1.71	27.13
3	24.77	5.06	23.75	17.87	1.15	17.50
4	22.47	5.10	20.00	15.43	0.80	15.31
5	14.00	4.15	15.00	3.72	0.44	3.66
6	64.68	12.40	65.00	49.45	3.65	50.00
7	37.96	3.96	38.00	29.99	1.77	30.16
8	28.20	8.35	26.00	24.46	3.73	25.30
9	24.23	5.99	25.00	18.57	2.16	19.12
10	21.38	5.30	20.00	18.06	1.43	18.22

[a] Five common procedures performed by conventional providers (lines 1–5) and five common procedures performed by CAM providers (lines 6–10).