

# Descriptive analyses of the integrity of a US Medicaid claims database

Sean Hennessy PharmD, PhD<sup>\*,1,2</sup>, Warren B. Bilker PhD<sup>1,2</sup>, Anita Weber PhD<sup>1</sup> and Brian L. Strom MD, MPH<sup>1,2</sup>

<sup>1</sup>Center for Clinical Epidemiology and Biostatistics, and Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA, USA

<sup>2</sup>University of Pennsylvania Center for Education and Research in Therapeutics (CERTs), Philadelphia, PA, USA

## SUMMARY

**Purpose** To examine the integrity of six Medicaid databases for use in pharmacoepidemiology research.

**Methods** We performed descriptive analyses to examine four categories of potential data errors: incomplete claims for certain time periods; absence of an accurate indicator of inpatient hospitalizations; missing hospitalizations for those aged 65 years and over; and diagnostic codes in demographic groups in which those conditions should be rare.

**Results** Prescription claims appeared to be missing intermittently in some states. No valid marker of inpatient hospitalizations could be found for three of six states. Hospitalizations appeared to be missing to varying degrees for those aged 65 years and over. Gross errors in diagnostic codes and demographic data did not appear to be widespread.

**Conclusions** Whenever possible, investigators using administrative data should perform macro-level descriptive analyses on the parent data set. In particular, researchers should examine the number of medical and pharmacy claims over time, looking for gaps. Validity of markers of hospitalization should be assessed. The accuracy of diagnosis and demographic data should be examined. Such a descriptive macro-level approach should be used to supplement, and perhaps precede validation of study outcomes using clinical records. Copyright © 2002 John Wiley & Sons, Ltd.

KEY WORDS — Medicaid; validity; pharmacoepidemiology; epidemiologic methods

## INTRODUCTION

US Medicaid claims data are often used for epidemiologic research. Although there are many advantages to their use, there are limitations as well. An important potential limitation is that the validity of diagnoses

included on medical claims is good for some conditions and poor for others.<sup>1</sup> In addition, because in the US, Medicare is almost always the primary payer for hospital claims for those aged 65 years and above, Medicaid hospitalization claims may be incomplete in this age group. Although this potential problem is recognized,<sup>1</sup> its magnitude has not been described. Other factors such as errors in the transfer of data could result in missing data that might go undetected.

We obtained access to Medicaid claims data from six states, described below. This allowed us to carry out to macro-level quality assurance checks of these six data sets. We examined evidence for four general categories of potential data errors: incomplete claims for certain time periods; absence of an accurate indicator to identify inpatient hospitalizations; incomplete hospitalization data for those aged 65 years and over;

\* Correspondence to: Sean Hennessy, PharmD, PhD, University of Pennsylvania School of Medicine, 803 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021, USA.  
E-mail: shenness@cceb.med.upenn.edu

Contract/grant sponsor: National Institute on Aging.

Contract/grant numbers: 1R01 AG14601.

Contract/grant sponsor: Agency for Healthcare Research and Quality.

Contract/grant numbers: 1 FS32 HS00066 and U18-HS10399.

Contract/grant sponsor: PharMark Corporation.

and diagnostic codes in demographic groups that would be expected to have a low frequency of those conditions.

## MATERIALS AND METHODS

### Data source

A data vendor that provides Medicaid programs with software and services related to the conduct of drug utilization review (DUR) programs<sup>2</sup> provided us with a reportedly complete set of medical and pharmacy claims for six Medicaid programs covering a defined period. Such vendors are a common source of Medicaid data, including the Computerized On-line Medical Pharmaceutical Analysis and Surveillance System (COMPASS).<sup>1</sup> The time period of available data and the annual number of enrollees in the Medicaid programs are listed in Table 1.

### Analyses

Analyses were descriptive in nature. We first explored the possibility that there were blocks of missing claims for certain time periods. In particular, we examined, on a monthly basis, the number of claims for dispensed drug prescriptions, and the number of enrollees with an outpatient medical claim, looking for apparent gaps.

We then examined the use of claims data to identify inpatient hospitalizations. In particular, we compared, for US federal fiscal year 1995 (1 October 1994 to 30 September 1995), the number of enrollees with a claim for an inpatient hospitalization with the number reported by the US Centers for Medicare & Medicaid Services (CMS; formerly the Health Care Financing Administration (HCFA)) as the HCFA 2082 report (<http://www.hcfa.gov/medicaid/msis/mstats.htm>, accessed 22 May 2001).

For each state, we next examined the degree of apparent incompleteness in hospitalization claims for

those aged 65 years and over. In particular, we examined the number of hospitalizations per enrollee, stratified by age group. Because the frequency of hospitalization among adults should increase monotonically with age, we looked for a negative inflection point at age 65 years as an indicator of incomplete hospitalization claims in this age group.

In order to examine the overall validity of diagnosis and demographic data, we first identified a number of disorders that would be expected to be present exclusively or predominantly in particular demographic groups: female-specific disorders in females, complications of childbirth and pregnancy in those aged 60 years and under, and lung cancer in those aged 40 years and over. The diagnostic codes used are listed in Appendix 1. We then visually examined, for each state for each month, the number of diagnoses in the expected demographic group versus the number in the unexpected group. We recognized in advance that some of the selected conditions would not occur exclusively in the expected group. However, if the data were fundamentally sound, then the number of disease-demographic 'matches' would be much greater than the number of 'mismatches'.

## RESULTS

Figure 1 presents the number of prescription drug claims per month in the three larger states (States A, B, and C), and Figure 2 presents these numbers for the three smaller states (States D, E, and F). Data from State C and, to lesser degrees, States E and F were reasonably level. State B was characterized by multiple downward spikes and a single upward spike above baseline, and State A was characterized by multiple upward spikes and a single downward spike from baseline. State D was characterized by an increasing trend over much of the observed time period.

Figures 3 and 4 present the number of outpatient medical claims per state per month. In general,

Table 1. Characteristics of study states

State	Expected data Start date	Expected data End date	Number of Medicaid recipients*				
			1992	1993	1994	1995	1996
A	July 1993	February 1996	2 557 701	2 742 494	2 907 963	3 035 477	3 281 016
B	May 1992	February 1996	1 313 140	1 395 566	1 441 034	1 551 949	1 454 152
C	August 1992	February 1996	554 477	609 386	668 765	695 458	636 176
D	November 1991	February 1996	360 039	386 531	390 628	393 613	358 121
E	April 1992	February 1996	60 186	89 041	96 206	98 708	101 271
F	December 1991	February 1996	42 401	46 262	50 554	51 374	51 231

\*Source: <http://www.hcfa.gov/medicaid/msis/mstats.htm> [accessed 22 May 2001].

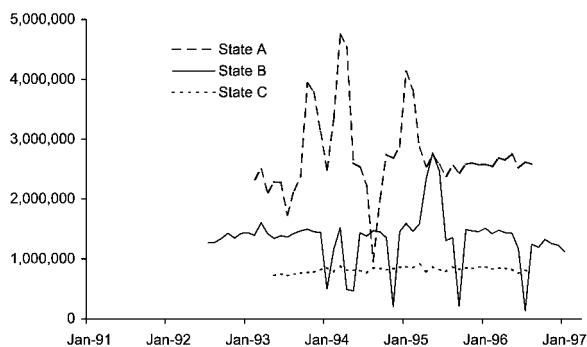


Figure 1. Prescription drug claims per month in the Medicaid programs of States A, B and C

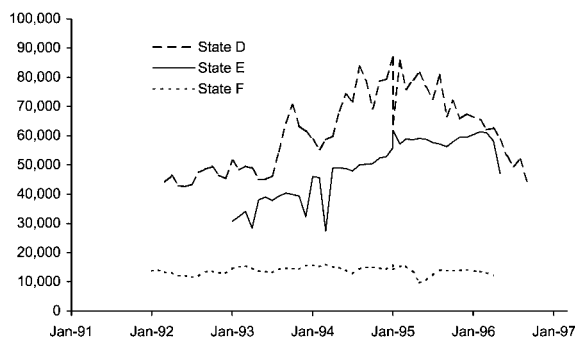


Figure 4. Number of enrollees per month with an outpatient medical claim in States D, E and F

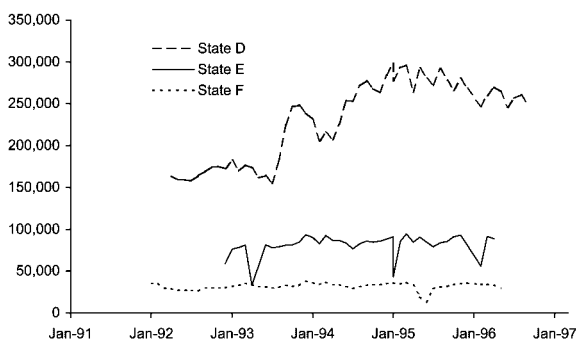


Figure 2. Prescription drug claims per month in the Medicaid programs of States D, E and F

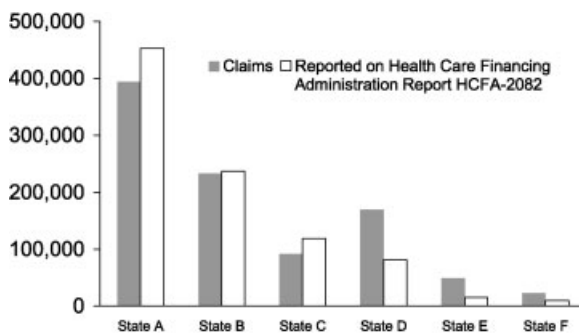


Figure 5. Medicaid enrollees with an inpatient hospitalization as determined by Medicaid claims data and as reported on the Health Care Financing Administration Report HCFA-2082, stratified by state

positive and negative spikes are less prominent for outpatient medical claims than for prescription drug claims.

Figure 5 compares the number of Medicaid enrollees with inpatient hospitalization in US federal fiscal year 1995 derived from claims data with the value obtained from aggregate data published by the CMS. In States A, B and C, the numbers derived from claims

were somewhat lower than those reported by the CMS, with the degree of underestimation varying from less than 2% in State B to 30% in State C. In States D, E and F, hospitalization claims greatly overestimated the figures reported by the HCFA, with the degree of overstatement ranging from 107% in State D to 214% in State E. In order to evaluate the possibility that a different data field should have been used to identify an inpatient admission in States D, E and F, we examined other all plausible fields that were provided to us. However, even working with the vendor, we were not unable to identify a marker that yielded numbers that were similar to those reported by the CMS.

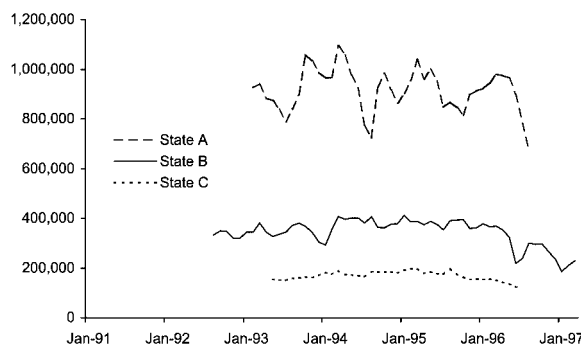


Figure 3. Number of enrollees per month with an outpatient medical claim in States A, B and C

Figure 6 presents, for each state, the cumulative number of inpatient hospitalizations per enrollee over the entire data period, stratified by age group. Because the amount of time represented varied by state, within-state rather than among-state comparisons are of interest. In all states, there were fewer hospitalizations per person in the 65–75-year-old age group than in the 45–64-year-old age group. The degree to which the apparent frequency of hospitalization dropped (where

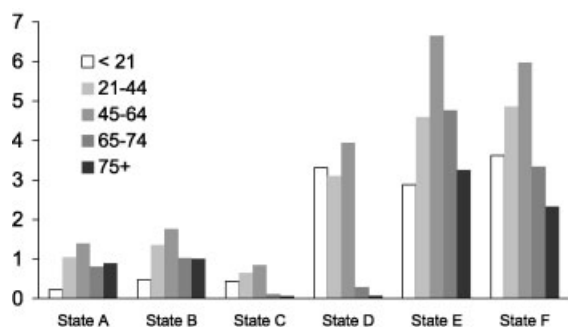


Figure 6. Ratio of number of Medicaid claims for an inpatient hospitalization, cumulative over all months, to Medicaid enrollee population size, stratified by age group

it should have increased) at age 65 years varied from 28% in State E to 93% in State D.

Figures 7 to 24 are state-specific plots of the number of enrollees with various conditions in the expected versus unexpected demographic group, per month. These data are plotted on a logarithmic scale to accommodate the large differences observed between expected and unexpected demographic groups. All of the plots show dramatically more disease-demographic matches than mismatches. There was an apparent peak in mismatches in State F in late 1993 to early 1994 (Figures 12 and 18).

DISCUSSION

*Implications of findings for the use of the current data set*

The constancy of number of prescription claims in State B, and to a lesser degree in States E and F, suggested that large blocks of prescription drug claims

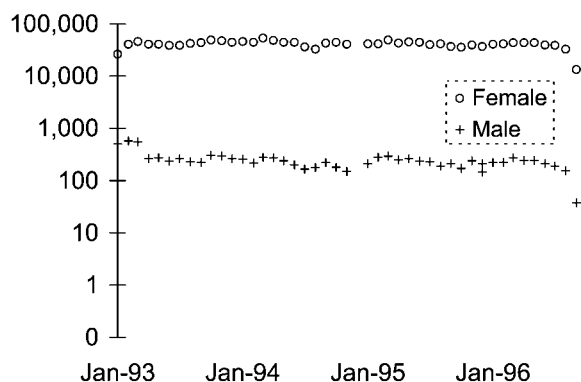


Figure 7. Number of enrollees with a female-specific diagnosis in the Medicaid program in State A, stratified by sex and time and plotted on a logarithmic scale

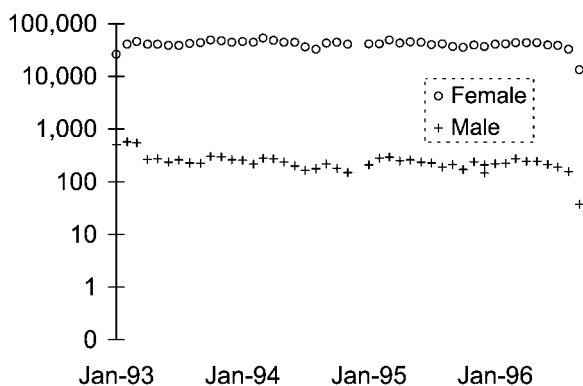


Figure 8. Number of enrollees with a female-specific diagnosis in the Medicaid program in State B, stratified by sex and time and plotted on a logarithmic scale

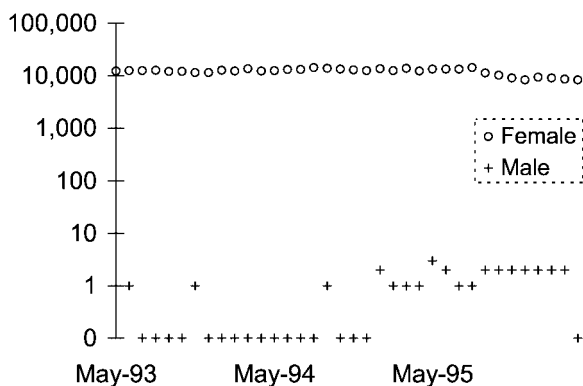


Figure 9. Number of enrollees with a female-specific diagnosis in the Medicaid program in State C, stratified by sex and time and plotted on a logarithmic scale

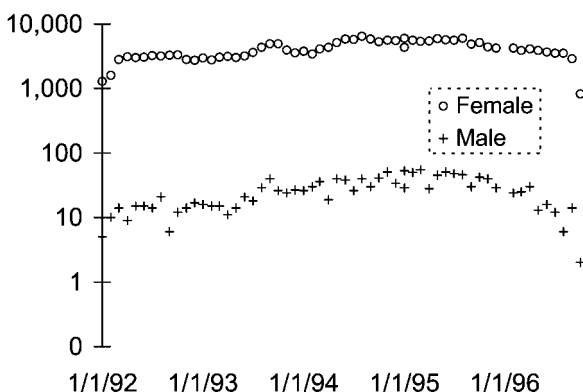


Figure 10. Number of enrollees with a female-specific diagnosis in the Medicaid program in State D, stratified by sex and time and plotted on a logarithmic scale

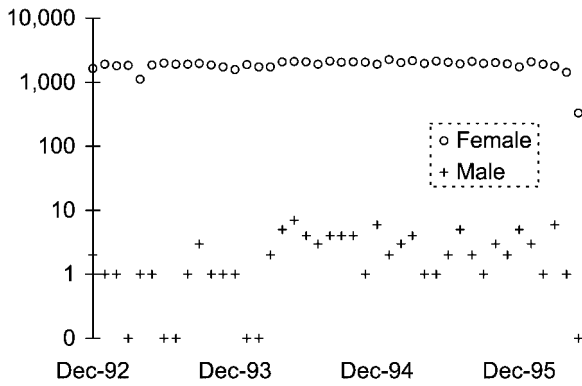


Figure 11. Number of enrollees with a female-specific diagnosis in the Medicaid program in State E, stratified by sex and time and plotted on a logarithmic scale

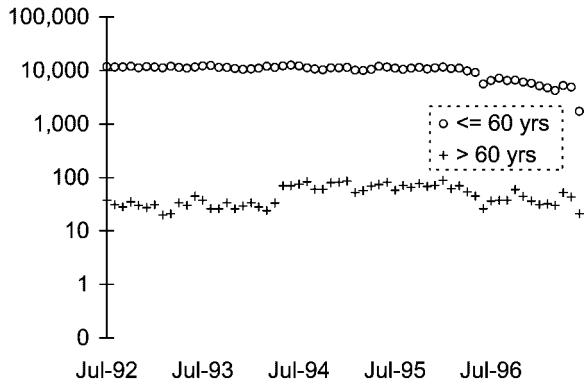


Figure 14. Number of enrollees with a diagnosis of complications of childbirth and pregnancy in the Medicaid program in State B, stratified by age and time and plotted on a logarithmic scale

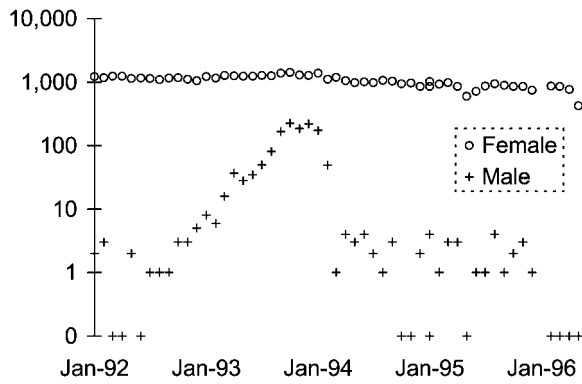


Figure 12. Number of enrollees with a female-specific diagnosis in the Medicaid program in State F, stratified by sex and time and plotted on a logarithmic scale

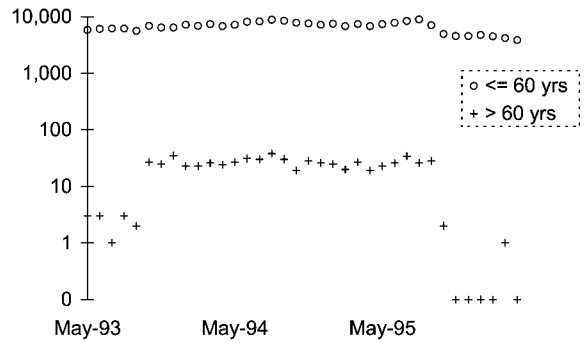


Figure 15. Number of enrollees with a diagnosis of complications of childbirth and pregnancy in the Medicaid program in State C, stratified by age and time and plotted on a logarithmic scale

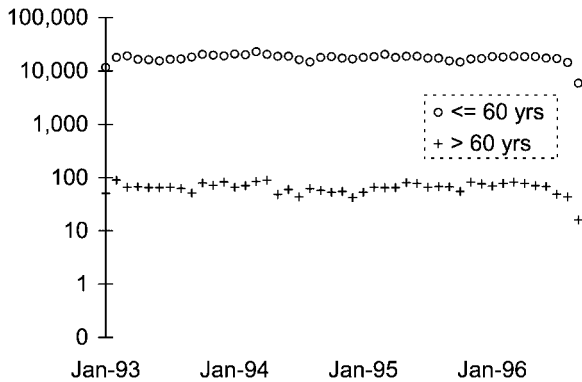


Figure 13. Number of enrollees with a diagnosis of complications of childbirth and pregnancy in the Medicaid program in State A, stratified by age and time and plotted on a logarithmic scale

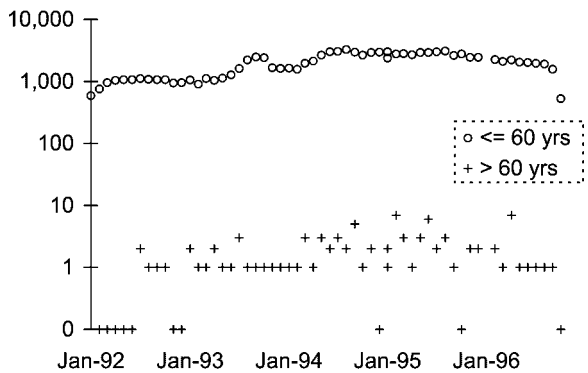


Figure 16. Number of enrollees with a diagnosis of complications of childbirth and pregnancy in the Medicaid program in State D, stratified by age and time and plotted on a logarithmic scale

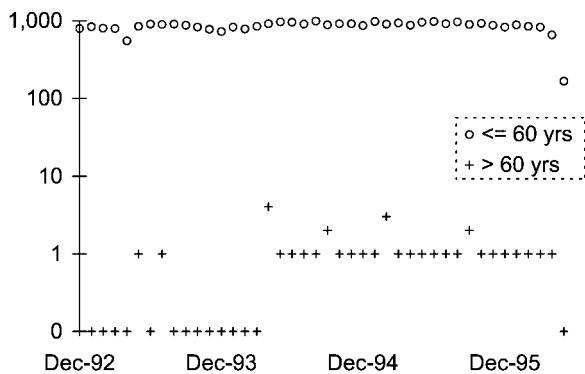


Figure 17. Number of enrollees with a diagnosis of complications of childbirth and pregnancy in the Medicaid program in State E, stratified by age and time and plotted on a logarithmic scale

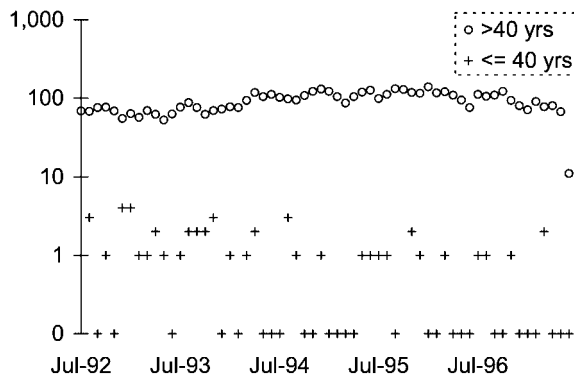


Figure 20. Number of enrollees with a diagnosis of lung cancer in the Medicaid program in State B, stratified by age and time and plotted on a logarithmic scale

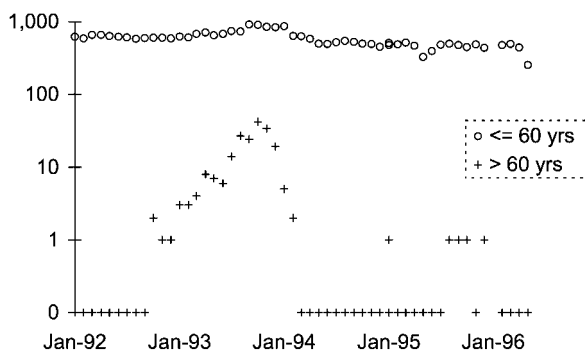


Figure 18. Number of enrollees with a diagnosis of complications of childbirth and pregnancy in the Medicaid program in State F, stratified by age and time and plotted on a logarithmic scale

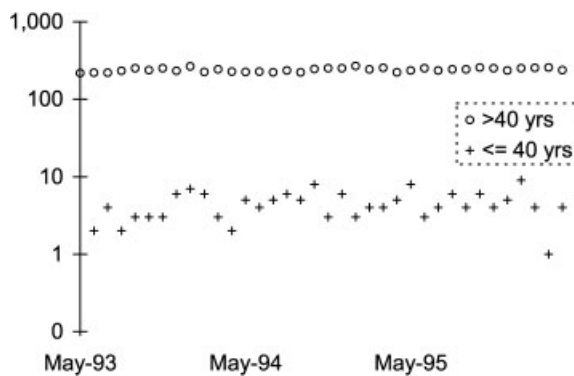


Figure 21. Number of enrollees with a diagnosis of lung cancer in the Medicaid program in State C, stratified by age and time and plotted on a logarithmic scale

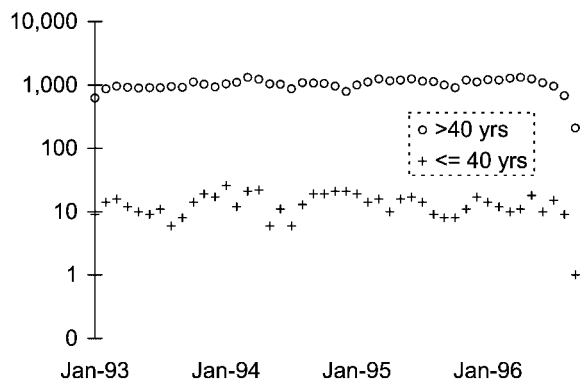


Figure 19. Number of enrollees with a diagnosis of lung cancer in the Medicaid program in State A, stratified by age and time and plotted on a logarithmic scale

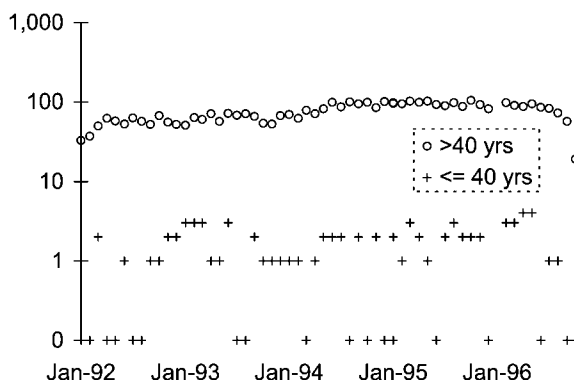


Figure 22. Number of enrollees with a diagnosis of lung cancer in the Medicaid program in State D, stratified by age and time and plotted on a logarithmic scale

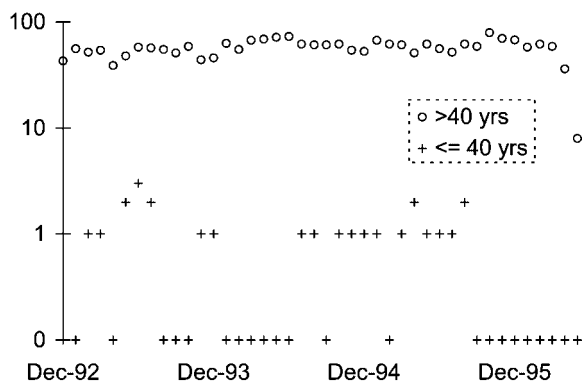


Figure 23. Number of enrollees with a diagnosis of lung cancer in the Medicaid program in State E, stratified by age and time and plotted on a logarithmic scale

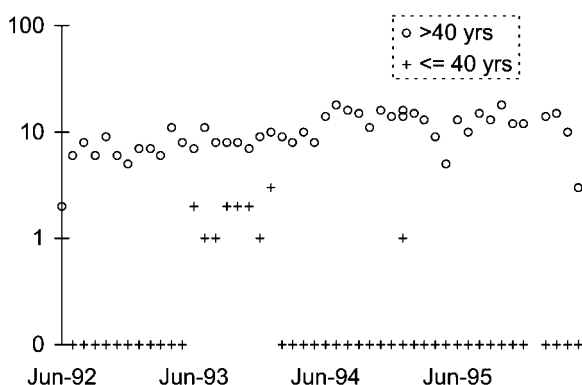


Figure 24. Number of enrollees with a diagnosis of lung cancer in the Medicaid program in State F, stratified by age and time and plotted on a logarithmic scale

are not missing in those data sets. Sporadic downward spikes, as observed in States A and B suggest that data may be intermittently missing in these states. The possibility that vendors might be missing prescription claims for particular time periods is supported by statements that appear in some annual reports of DUR programs. The multiple upward spikes in prescription claims seen in State A, and the single upward spike seen in State B are difficult to interpret without additional information. Whether these spikes represent duplicate claims, some other artifact, or are true peaks in prescription activity remains unknown. Also unknown is the source of the increasing trend in the number of prescription claims seen in State E (Figure 2). This increase was not accompanied by a concomitant increase in the number of Medicaid recipients in that state (Table 1).

The relative constancy of outpatient medical encounters over time does not suggest obvious large

gaps for this type of claim (Figures 3 and 4). However, we had no external source of numbers of outpatient medical claims to serve as the basis for comparison. Therefore, conclusions about the completeness of outpatient claims must be tentative.

We were unable to identify an apparently valid marker of inpatient hospitalizations for three of the six states examined. This makes studying the frequency of hospitalization possible in only half of the study states. Fortunately, the three states with an apparently valid hospitalization marker are the three largest states in this group.

As expected, our results indicated that hospitalization claims were incomplete for those aged 65 years and over. Given that we had no valid marker of hospitalization in States D, E and F, the age-specific number of hospitalizations in these states is uninterpretable. In States A, B and C, the degree of incompleteness was substantial, although it varied by degree. Because Medicare is usually the primary payer in the US for hospitalizations among the elderly, this is not an error of the Medicaid data *per se*, but rather a recognized problem in its use to study hospitalizations among the elderly. Clearly, inferences regarding hospitalizations of the elderly can be made only very cautiously if at all, using these data. Depending on how the data were to be used, this misclassification might well be non-differential with regard to exposure. When this is the case, effect measures such as relative risks would be biased toward the null.

Plots of the frequency of diagnoses in an expected versus unexpected demographic group suggested that gross errors in diagnostic codes and demographic data were not widespread. However, the results did suggest the presence of random coding errors in either diagnoses, or in demographic characteristics such as gender or date of birth. Without additional information we cannot discern whether the errors are in the disease codes or demographic data. The apparent peak in mismatches in State F in 1993–1994 was interesting to observe, but difficult to interpret without additional data. As expected, plots of disorders with less specificity for a particular demographic characteristic (e.g. lung cancer in those aged 40 years and over) showed smaller differences, than those with more specificity (e.g. gender-specific conditions).

*Implications for researchers using other administrative data*

Our results suggest that, whenever possible, investigators using administrative data should carry out macro-level descriptive analyses on the parent data set. In

particular, researchers should examine the number of claims of different types (e.g. prescription, inpatient medical, outpatient medical) over time, looking for apparent gaps. Validity of markers of hospitalization should be assessed, with comparison with an external standard undertaken whenever possible. The accuracy of diagnosis and demographic data should be assessed by examining the frequency of select diagnoses stratified by demographic group. Naturally, the examples presented in this paper are not exhaustive. Such a descriptive macro-level approach should ideally be used to supplement, and perhaps precede the practice of validating outcomes by examining clinical records, where these are available.

Investigators using administrative data to conduct epidemiologic research often obtain data only on the subset of enrollees who will be included in the results of the study. This precludes many of the macro-level quality assurance checks which should be carried out here. Nevertheless, our results point to some specific concerns that should be kept in mind by investigators who do not have access to the parent data set.

Investigators should consider the possibility of incomplete prescription claims in the execution and interpretation of studies. For example, longitudinal studies examining prescription refill patterns might incorrectly interpret incomplete claims as the failure of subjects to obtain prescription fills. Examining whether a greater than expected number of such gaps occur simultaneously in calendar time within the study cohort may help identify such gaps. The effect of these gaps on studies that follow patients for a given period after a prescription is filled would be to reduce the number of observations. However, it should not result in bias in either the absolute or relative risk of study outcomes. Studies adjusting for past receipt of chronically administered drugs should not be severely affected as long as such gaps were relatively rare. This is because individuals receiving the drug chronically will be likely to have also received the drug during a period in which prescription data are complete. Studies using receipt of a drug as the study outcome may miss outcomes. For studies using a control group, this incompleteness would most likely be non-differential with regard to exposure, and thus result in bias toward the null.

The validity of hospitalization markers is often crucial in defining study outcomes, and may be suspect. Investigators using hospitalization as part of a study outcome need to assure the validity of the available marker for this outcome.

As expected, hospitalization data for those aged 65 years and over was incomplete. The degree of incom-

pleteness was substantial in all states examined, and dramatic in some. Studies using Medicaid data linked to Medicare data would hopefully avoid this problem.<sup>3,4</sup> Medicaid studies without access to Medicare data may have limited ability to study hospitalization outcomes in the elderly. At the very least, in order to avoid data incompleteness as a source of information bias, investigators using Medicaid data should look for confounding and effect measure modification by age, with particular attention to the 65-year age threshold.

In conclusion, we undertook macro-level quality assurance checks on a typical source of Medicaid used for epidemiologic studies. These analyses were highly informative regarding the utility of these specific data sets. Whenever possible, investigators should carry out such macro-level descriptive analyses. When the parent data set is unavailable, additional care is needed in the analysis and interpretation of studies.

#### KEY POINTS

- Researchers using claims data rarely have the opportunity to carry out quality assurance of the whole data set
- Performing such analyses can reveal important limitations of the data
- Whenever possible, researchers should examine the 'parent' data set for apparent irregularities

#### ACKNOWLEDGEMENTS

We thank Jesse A. Berlin, Jeffrey L. Carson, Sandra A. Norman, Wendy P. Stephenson, and Stephen B. Soumerai for critiquing drafts of this paper, Stephen Durboro, Maureen Moffett, M. Lee Morse, Aida LeRoy, Joann Trianfo-Lincoln, Vanessa Hatch, Haik Novshadian, Ming Tang, Glen Hallums, Peter Hoffman, and Kevin Hall for technical support, and Sherri Stubblefield for secretarial assistance.

This study was funded by the National Institute on Aging (1R01 AG14601), the Agency for Healthcare Research and Quality (1 FS32 HS00066 and U18-HS10399), and the PharMark Corporation.

#### REFERENCES

1. Carson JL, Strom BL. Medicaid Databases. In *Pharmacoepidemiology* (3rd edn), Strom BL (ed.). John Wiley & Sons Inc.: Chichester, UK, 2000.
2. Hennessy S, Strom BL, Lipton HL, Soumerai SB. Drug utilization review. In *Pharmacoepidemiology* (3rd edn), Strom BL

- (ed.). John Wiley and Sons, Inc.: Chichester, UK, 2000; 505–524.
3. Shorr RI, Ray WA, Daugherty JR, Griffin MR. Incidence and risk factors for serious hypoglycemia in older persons using insulin or sulfonylureas. *Arch Intern Med* 1997; **157**: 1681–1686.
  4. Wang PS, Walker A, Tsuang M, Orav EJ, Levin R, Avorn J. Strategies for improving comorbidity measures based on Medicare and Medicaid claims data. *J Clin Epidemiol* 2000; **53**: 571–578.

## APPENDIX 1: ICD-9 CODES USED FOR QUALITY ASSURANCE ANALYSES

---

### Female-specific disorders:

Inflammatory diseases of female pelvic organs and other disorders of the female genital tract (614–616; 617–629 and subcodes); complications of pregnancy, childbirth and the puerperium (630–676 and subcodes);

malignant neoplasms of female genital organs (179, 180, 181, 182, 183, 184); benign neoplasms of female genital organs (219, 220, 221);

malignant neoplasms of the female breast (174 and subcodes);

poisoning due to contraceptives (962.2).

Complications of childbirth and pregnancy: 630 to 676 and subcodes

Lung cancer: 162

---