Linking utilization data with resident assessment data is an economical way to describe the characteristics and outcomes of nursing home residents. Administrative records of insurance claims activity have provided resource utilization data to researchers for some time. The development of the national Minimum Data Set (MDS) for use in long-term care facilities now routinely generates a standard set of data elements that previously were available only through chart reviews, if at all. We describe a project in which we linked resident assessment data with the corresponding Medicaid claims files. The merged data set contained information about the use of medical services and health status of this population. We describe the linkage process and the characteristics of the combined data — its strengths, limitations, and potential uses.

The study also provided an opportunity to compare Medicaid beneficiaries with the overall population of residents. We describe the differences between Medicaid and non-Medicaid residents on medical and functional characteristics, and thus specify the conditions under which it is appropriate to generalize from Medicaid residents to the overall nursing home population.

Combining Medicaid claims and resident assessment data presents an opportunity to create a longitudinal data set concerning the use of medical resources with assessment status of nursing home residents. We advocate expanded use of linked data sets for research on outcomes of care in the institutionalized elderly. In this article, we present some of the methodological and social concerns that must be resolved before these linkages can be used routinely and with confidence.

Background

Information from claims generated by Medicaid Assistance Programs (Medicaid) has been particularly useful as a source of secondary data about health care resource use. Medicaid data, collected unobtrusively, provide a constant stream of information on medical resources used by a large number of patients. The data are available in a timely fashion and at relatively low cost (Bright, Avorn, & Everitt, 1989; Paul, Weis, & Epstein, 1993; Romano & Luft, 1992). The limitations of Medicaid and other administrative claims data for research purposes is well described in the literature (Bright et al., 1989; Romano & Luft, 1992; Stregachis, 1988). The lack of diagnostic data is a major factor limiting research because a complete medical description is not necessary for physicians or institutions to receive payment for services. Diagnostic data often are inaccurate as well as incomplete (Bright et al., 1989). Furthermore, claims records do not include information about social history, and there may be gaps in the record as persons gain and lose beneficiary status. A problem peculiar to elderly persons is that claims data may be stored in multiple files when individuals are eligible to receive care under both Medicaid and Medicare. Finally, there is concern about using the data for clinical studies when the data were submitted for reimbursement purposes. Claims may be prepared by clerical staff and could be biased by financial considerations inherent in the payment scheme.

Key Words: Minimum data set (MDS), Medicaid, Data linkage
Resident assessment data required for multiple purposes also has its advantages and disadvantages. One requirement for a standardized set of patient information emanates from the Federal Government, which has sponsored the development of a Minimum Data Set (MDS) as a mechanism for implementing and improving the quality of care of residents in long-term care settings (Morris et al., 1990). Although MDS data are filed in the patient medical record rather than in a central data base, the MDS defines the content and format of the data that are collected. There is a requirement to submit patient-level data to a centralized location in some states, which reimburse nursing homes on the basis of case mix, that is, where payment varies according to resident needs and nursing resource use (Arling, Zimmerman, & Updike, 1989). The data elements of the MDS and case mix reimbursement initiatives are similar. They include measures of impairment in activities of daily living (especially eating, incontinence, transferring, and ambulation), and clinical conditions requiring specialized nursing care, intensive rehabilitation, or behavioral management.

Standardized resident assessment data are seen by some as a major new resource for researchers interested in long-term care. Others have reservations about the utility of MDS data, particularly when the purpose of collecting the information is motivated by clinical or financial considerations rather than by the demands of the research (Hawes, Phillips, Mor, Fries, & Morris, 1992; Phillips, Chu, Morris, & Hawes, 1993a; Teresi & Holmes, 1992). It is doubtful that a single measure, designed to tap care-relevant aspects of a resident's status and functioning, will address aspects of interest in the depth sufficient for research purposes. There are concerns about the lack of standardization for individual data elements, due in part to the inter-rater reliability arising from the use of multiple, largely untrained data collectors across sites, and in part to inconsistencies in sources of information within a single site over time. Others fear that tying reimbursement to patient status introduces biases in the data toward maximizing payment rather than providing a comprehensive description of patient status. On the other hand, the MDS data already are being used to inform public policy debates (Arling, Rhyther, Collins, & Zimmerman, 1991; Philips, Hawes, & Fries, 1993b; Riter & Fries, 1992), and are seen as an untapped resource for improving quality and outcomes in the nursing home setting (Kane, 1995).

Table 1 is a comparative listing of claims elements included in the Medicaid Management Information System and data collected as part of the MDS.

### Methods

In the spring of 1985, the Wisconsin Division of Health performed resident assessments that included data elements similar to the national MDS. The Resident Evaluation Survey (RES) used in this pilot study was designed to capture the presence and severity of clinical conditions and services most commonly encountered in long-term care facilities and those which require the most intensive use of human resources. The development and testing of the RES instrument, including support for its reliability and validity, has been described in detail elsewhere (Arling et al., 1989).

In 1987, all Wisconsin nursing homes (N = 453) were given the option of supplying individual resident assessments in place of aggregate data for their annual utilization report to the State. This option was selected by 123 facilities, with a total population of 14,917 residents. Their assessments covered all residents, both Medicaid and non-Medicaid, and refer to the resident's condition on December 31, 1987. Medicaid data for our study were obtained from the state’s fiscal intermediary and RES data were supplied by the Wisconsin Center for Health Statistics, with approval of the Division of Health. All analyses were performed at the Center for Health Systems Research and Analysis on a VAX 780. Statistical analyses were conducted using SPSS-X software, release 3.0.

To check the quality of RES data, we examined the following: (a) the number of assessments per home compared to their reported census; (b) the number of the residents classified as Medicaid in the RES data for which no valid identification (ID) number could be located; (c) the proportion of missing data points on any of the items. No remarkable discrepancies were found. Additionally, the data were reasonably

### Table 1. Comparison of Data Elements Present in Medicaid Management Information Systems and Collected as Part of the National Minimum Data Set (MDS)

<table>
<thead>
<tr>
<th>Data Element</th>
<th>Medicaid</th>
<th>MDS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Date of birth</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Reimbursement source(s)</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Date of admission to long-term care facility (LTCF); level of care; developmental disability; guardian; protective placement, involuntary commitment, locked ward</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Mental health history; history of alcohol or other drug abuse</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Disease diagnoses</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Medical services: date(s), amount, description, amount paid, provider</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Hospital</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Nursing home</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Physician</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Pharmaceuticals</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Functional status: ambulation, transfers</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>nutrition, hygiene</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Incontinence</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Sensory impairment: hearing, vision</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Cognitive status; memory impairment</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Behavior: verbally or physically abusive, suicide attempts, activity patterns</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Use of antipsychotic/depressant medication</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Special services: tracheostomy, ostomy, appliances, intravenous medication, skin care, specialized nursing</td>
<td>**</td>
<td>**</td>
</tr>
<tr>
<td>Physical/occupational therapy</td>
<td>**</td>
<td>**</td>
</tr>
</tbody>
</table>
complete. There were 125 (1.2%) persons reported eligible for Medicaid for which no valid ID was given. There were no missing responses on the health and functional measures and treatment items. Among the demographic elements, item non-response ranged from .2 percent on age and gender, to 1.2 percent on reimbursement and level of care.

We analyzed the data for this study in three stages. First, the distribution of the voluntarily submitted RES data was compared to state-published population data as a check of whether the respondents in the data set adequately represented the state nursing home population (Stergachis, 1988). Second, comparisons between Medicaid and non-Medicaid residents were made across RES data elements concerning residents’ status (psychosocial problems, sensory impairments, nursing treatments, and medical problem indicators), assistance with activities of daily living (ADLs), and demographic and other characteristics. This procedure was conducted to determine the potential for bias created by basing research upon a data linkage where only Medicaid recipients would represent the overall population of nursing home residents. Two-tailed t-tests with separate variance estimates were performed to examine differences between the Medicaid and non-Medicaid residents in the RES pool.

In the third and final step, unique patient identification numbers from RES records were matched against Medicaid claims filed for services provided during December 1987, and paid before June 1, 1988. For every resident match, variables common to both data sets were compared, including gender, age and use of antipsychotic or antidepressant medication. Finally, ICD-9-CM diagnoses reported in the RES data set were matched against ICD-9-CM diagnoses supplied on nursing home, hospital, and physician claims.

Results

Although RES data were not obtained from a randomly selected sample, both the participating nursing homes and the residents were representative of the overall population with respect to age, gender, method of reimbursement, level of care, length of stay, nursing home bed capacity, and region of the state. The RES sample compared favorably on 20 mental and physical status indicators for which population data were available. The sole difference between the sample and the overall population was an under-representation of proprietary homes in the RES sample. Although 45% of the state’s nursing homes are proprietary, they comprised only 23% of Medicaid residents. Consequently, government and nonprofit facilities accounted for 37% and 39% of the RES sample, respectively, whereas they make up 23% and 32% of the total number of homes in the state.

Next, we compared the RES data responses between Medicaid and non-Medicaid residents. The comparison yielded statistically significant differences on 8 of 17 problem indicators and 2 of 4 ADL items, as shown in Table 2. Medicaid residents had fewer physical and functional impairments, were more likely to be cognitively impaired, and exhibited a greater degree of undesirable behaviors. Medicaid recipients on average were 7.9 years younger than non-Medicaid residents, resided in the nursing home for 40.1 months longer, and were more apt to be unmarried.

However, we also observed that 97.2% of the 1,434 residents with developmental disabilities, and 95.7% of the 1,479 persons under the age of 55 were Medicaid recipients. When these groups were dropped from the analysis, the differences between the Medicaid and non-Medicaid residents were not as pronounced (see Table 2). There was no longer a significant difference in the problems of mobility and incontinence, nor in the prevalence of hearing impairments. Medicaid recipients received either an equal or greater amount of assistance with ADLs compared to non-Medicaid residents. Again, the presence and severity of psychosocial problems were greater among Medicaid recipients and the prevalence of memory impairment did increase slightly. The differences in average age, 1.9 years, and length of stay, 33.6 months, were reduced but remained significantly different.

For the linkage portion of the study, there were 10,416 residents in the RES data set with a valid Medicaid ID number. When the ID numbers were matched to ID numbers on nursing home claims, a paid claim was found for 9,063 (87.0%). At least three reasons partially account for the number of residents who could not be paired with a paid claim. On January 1, 1988, the state assigned new patient IDs. Failure to correctly record the new numbers reduced the number of successful linkages. Second, despite the six-month lag, about 5% of the Medicaid nursing home claims were outstanding. Third, data collection errors were possible due to inexperience with RES forms by nursing home personnel and the unexpected high proportion of manual data entry by state workers. Although nursing homes were allowed to submit data on computer disks provided by the state, few used this option.

Within the group of 9,063 persons with matched ID numbers, gender and year of birth agreed with Medicaid records in 99.3% and 98.2% of cases, respectively. The only medication use information requested on the RES was a single item to indicate whether the patient received antidepressant or antipsychotic drug therapy. According to the RES responses, 4,236 (46.7%) of the linked subjects received antidepressant or antipsychotic drug therapy. Of these 2,942 resided in a nursing home for which drugs are reimbursed on a fee-for-service basis. The transition to the new claims processing system and newly assigned patient IDs were a particular problem with pharmacy claims, resulting in processing delays. Approximately 25% of the total number of pharmacy claims anticipated were not on file at the time of the analysis. Nonetheless, a paid claim for an antidepressant or antipsychotic drug was located for 2,300 persons (78.2%).

Vol. 36, No. 4, 1996
For the RES, caregivers were asked to list the ICD-9-CM code for as many as four disabling conditions present on the assessment date. There were a total of 33,937 diagnostic codes provided from RES data, an average of 3.74 diagnostic codes per resident. Of these, 12,983 (38.3%) were matched to a diagnosis present in claims data at the three-digit level. There were 21,315 unique diagnoses obtained from Medicaid claims. About one half (49.6%) of the total. The figure also depicts the amount of information available from the RES data relative to the amount obtained from claims. More than 80% of the total diagnostic codes were present in RES data, whereas about one half were obtained from Medicaid claims.

Table 3 shows the number of diagnoses from each data source, matched and unmatched, for selected diagnostic categories. For all except two categories, RES data contained more diagnoses than the claims data. Of the diagnostic groups investigated, only claims with a diagnosis of traumatic injury were well matched to the same number of that diagnosis in the RES data. The proportion of claims which overlapped between the data sets ranged from 9.5% for constipation to 53.4% for psychoses. The highest proportion of overlap occurred among the mental and nervous system disorders of all types.

For the RES, caregivers were asked to list the ICD-9-CM code for as many as four disabling conditions present on the assessment date. There were a total of 33,937 diagnostic codes provided from RES data, an average of 3.74 diagnostic codes per resident. Of these, 12,983 (38.3%) were matched to a diagnosis present in claims data at the three-digit level. There were 21,315 unique diagnoses obtained from Medicaid claims. About one half (49.6%) of the total. The figure also depicts the amount of information available from the RES data relative to the amount obtained from claims. More than 80% of the total diagnostic codes were present in RES data, whereas about one half were obtained from Medicaid claims.

Table 3 shows the number of diagnoses from each data source, matched and unmatched, for selected diagnostic categories. For all except two categories, RES data contained more diagnoses than the claims data. Of the diagnostic groups investigated, only claims with a diagnosis of traumatic injury were well matched to the same number of that diagnosis in the RES data. The proportion of claims which overlapped between the data sets ranged from 9.5% for constipation to 53.4% for psychoses. The highest proportion of overlap occurred among the mental and nervous system disorders of all types.

Although some of the lowest percentage of overlap occurred for acute con-
conditions such as peptic ulcer disease (19.1%) and urinary tract infection (11.8%), the percentage of matched diagnoses for other acute conditions such as fractured femur (32.8%) was higher. We found no consistent pattern to diagnoses that were reported in either data source.

Discussion

Data linkages that build comprehensive records of patients' experiences in nursing homes and health states across time and site fill an important gap. They provide a way to study the relationships between process and outcomes in an important but neglected patient population (Aronow & Coltin, 1993; Kane, 1995). Studies using data linkages of the type described here are economically efficient because Medicaid data are merged with information previously available only through costly and time-consuming chart review or clinical observation. The alternative of medical records extraction is not only laborious but subject to human error and limited by the accuracy and completeness of the documentation. Either data source in isolation is far less useful for outcomes-oriented studies. Whereas patient assessments provide information only available by observing patients, the observations are recorded at defined intervals. Claims data are an important adjunct to assessment data since they provide continuous information about the medical care process and resident outcomes that occur during the interval between assessments (Fogel & Mor, 1994). The utility of existent data is subject to resolution of methodological issues and ethical concerns.

Our experience in drug use studies illustrates the benefit of linking assessment and Medicaid data sets for research. Medicaid data are particularly useful for research on drug use among elderly persons residing in nursing homes. Unlike the younger, ambulatory Medicaid population, gaps in the claims history of nursing home residents due to loss of coverage are unusual. Furthermore, it is relatively safe to assume that the patients consume the prescription medication dispensed. Nearly two thirds of the pharmacists supplying medication to nursing homes in Wisconsin use a unit dose distribution system and submit claims only for medication actually consumed (Lipowski,
While conclusions and implications based upon one would add marginally to the total number of diagnosis of alcohol or other drug dependence whenever the RES data recorded a history of alcohol or drug abuse in one of two places, by diagnostic code or direct questioning (Collins & Bigelow, 1988). Routine availability of MDS data elements would expand the opportunity to examine a treatment modality, such as drug use, with the continuous record of other health and social services documented in Medicaid claims, and observe the resultant social, clinical, and fiscal outcomes.

Though the data set described in this study is of limited current interest, it provided the opportunity to test the feasibility and characteristics of merging Medicaid and assessment data. First, having both Medicaid and non-Medicaid recipients within the RES data provided information about how the Medicaid population compares to all persons residing in nursing home settings. We found some differences between Medicaid and private pay residents in a pattern consistent with previous research from Colorado (Shaughnessy, Schlenker, & Polesovsky, 1986). The types and degree of differences may vary in other states, however, since Medicaid programs and medical care utilization are not uniform among nursing home populations in different states (Ray, Federsepp, Baugh, & Dodds, 1987). Data such as ours could assist researchers in extrapolating from Medicaid data to non-Medicaid populations through statistical modeling techniques.

We acknowledge two potential limitations to the data presented here. The first limitation concerns conclusions based upon a series of bivariate tests, for as the number of variables being compared increases, the likelihood of a Type I error increases (Shaughnessy et al., 1986). Thus, greater confidence is placed upon general trends and findings that are consistent across analyses and subpopulations, while conclusions and implications based upon one or two variables in a profile are to be avoided. The results of this study, however, are consistent with previous research (Bright et al., 1989; Jollis et al., 1993; Paul et al., 1993; Romano & Luft, 1992; Roos, Sharpe, & Cohen, 1991). The second limitation is due to the large sample size. Differences that are statistically significant may not be clinically significant. These results do provide information to others about specific variables that may be relevant to studies involving Medicaid populations.

Compared to Medicaid claims, diagnostic data provided on RES forms were more complete. Although some of the unmatched ICD-9-CM codes may be the result of unpaid claims, those claims would add marginally to the total number of diagnoses available for comparison. About one half of the eligible hospital claims were outstanding at the time of the analysis because hospital claims are often submitted in sequence to multiple parties, with Medicaid being the payor of last resort. However, fewer than 2% of aged Medicaid recipients residing in a nursing home receive services from a hospital in any given month, and the additional claims are not likely to contribute substantial new information, due to the pervasive problem of limited diagnostic coding.

In addition to problems related to claims processing, procedural changes, and delays, there is a fear that administrative data could also be potentially biased. Although tying resident assessments to reimbursement introduces the potential for a bias in recording the data, other forces discourage the introduction of systematic bias. Resident assessments are monitored during annual inspections of care. Not only must the assessment records themselves be accurate, but the plan of care for the patient must be appropriate and reflect the findings. Medicaid providers also are subject to financial audit. Organizations with a large and significant proportion of their income derived from a single payor anticipate close scrutiny and the serious consequences of questionable data found by auditors.

Prospects for the increased availability of nursing home resident data bases for research are promising but not secured. Politics and the outcomes of public policy debates will affect the content and availability to researchers (Bright et al., 1989; Paul et al., 1993; Romano & Luft, 1992). The Omnibus Budget Reconciliation Act of 1987 (OBRA '87) mandated the implementation of standardized nursing home resident assessments in support of the call for an outcome orientated inspection system (Health Care Financing Administration, 1988). The Health Care Financing Administration (HCFA) selected five states to participate in a large scale demonstration project focusing on the integration of case-mix payment and quality assurance systems in nursing homes. These data have been aggregated and represent an opportunity to test the utility of this approach. As noted earlier, reports of the project and research using those data are now beginning to appear in the medical literature and to form public policy debate (Arling et al., 1991; Phillips et al., 1993b; Riter & Fries, 1992). HCFA has stated its commitment “to the development of a national data base of this type as the end product of our efforts to collect standardized assessment information on all residents of certified nursing homes” (Health Care Financing Administration, 1988).

Even if national, standardized patient data bases are compiled, they may not be made available for research until questions of “...privacy, authorized access to, and uses of the data, and a number of other technically and legally complex and politically sensitive matters” are resolved (Institute of Medicine, 1986). The creation of comprehensive centralized computer files is likely to raise concerns regarding patients’ right to privacy and industry rights to proprietary information. However, there are those who would maintain that electronic storage of medical records can be even more secure than paper-based records (Drislane, 1994). Concerns about the use and disclosure of health data and issues related to privacy, confidentiality, and security of centralized data...
were the subject of extensive study by the Institute of Medicine (IOM). The report envisioned the establishment of regional health data organizations, which would monitor the uses and users of data to the benefit of the public while protecting individual providers and patients from any harmful or unfair consequences of use (Institute of Medicine, 1994).

In conclusion, data similar to the national MDS were linked with Medicaid-paid claims. This data linkage provided an opportunity to compare data elements common to both data sets and to explore the potential for using Medicaid data to represent the nursing home population. Identifying information was substantially congruent, but diagnostic data only partially overlapped. Data linkages of this type provide a potential resource for research and system management. Investigators should keep abreast of developments of national data bases under health care reform initiatives to ensure that essential data elements are included and that the data are accessible for the purposes of researchers.

References


Received December 7, 1994
Accepted February 7, 1996

Vol. 36, No. 4, 1996 447