0885-713X/95/1004–0162$03.00/0
AMERICAN JOURNAL OF MEDICAL QUALITY

Vol. 10, No. 4
Winter 1995

Developing a Quality Improvement Database Using Health Insurance Data: A Guided Tour with Application to Medicare's National Claims History File

Stephen T. Parente, Ph.D., M.P.H.,*† Jonathan P. Weiner, Dr.P.H.,‡ Deborah W. Garnick, Sc.D.,§ Thomas M. Richards, M.S.E.E.,‡ Jinnet Fowles, Ph.D.,¶ Ann G. Lawthers, Sc.D.,‖ Paul Chandler, B.A.,‡ and R. Heather Palmer, M.B., B.Ch., S.M.‖

*Center for Health Affairs, Project HOPE, Bethesda, Maryland, ‡Health Services Research and Development Center, The Johns Hopkins University School of Public Health, The Johns Hopkins University, Baltimore, Maryland, §Institute for Health Policy, Heller Graduate School, Brandeis University, Waltham Massachusetts, ¶Park Nicollet Medical Foundation, Minneapolis, Minnesota, and ‖Center for Quality of Care Research and Education, Harvard School of Public Health, Boston, Massachusetts

Health policy researchers are increasingly turning to insurance claims to provide timely information on cost, utilization, and quality trends in health care markets. This research offers an in-depth description of how to systematically transform raw inpatient and ambulatory claims data into useful information for health care management and research using the Health Care Financing Administration's National Claims History file as an example. The topics covered include: (a) understanding the contents and architecture of claims data, (b) creating analytic files from raw claims, (c) technical innovations for health policy studies, (d) assessing data accuracy, (d) the costs of using claims data, and (e) ensuring confidentiality. In summary, claims data are found to have great potential for quality of care analysis. As in any analysis, careful development of a database is required for scientific research. The methods outlined in this study offer health data novices as well as experienced analysts a series of strategies to maximize the value of claims data for health policy analysis.

Public and private agencies are increasingly turning to insurance claims to provide an overview of cost, utilization, and quality trends in health care markets. One of the richest sources of such data is the Health Care Financing Administration (HCFA). HCFA has recently created a new database encompassing all of Medicare's ambulatory and inpatient claims associated with its 35 million beneficiaries[1] (21). It is called the National Claims History (NCH) file. Medicare data are a huge resource only now beginning to be tapped for medical effectiveness, cost, and quality analyses. A thorough understanding of the Medicare data is crucial both for analysts who wish to use the data and for policy makers who need to take action based on those analyses.

† To whom correspondence should be addressed at Project HOPE Center for Health Affairs, 7500 Old Georgetown Road, Suite 600, Bethesda, MD 20814.

Supported by the Health Care Financing Administration, Health Standards and Quality Bureau as part of the HCFA's Project to Develop and Evaluate Methods to Promote Ambulatory Care Quality (DEMPAQ) under subcontract to the Delmarva Foundation for Medical Care (PRO Contract Modification: 500-89-0624).

Dr. Parente is Senior Research Director, Project HOPE Center for Health Affairs and Associate, The Johns Hopkins University Department of Health Policy and Management. Dr. Weiner is Professor and Deputy Director, The Johns Hopkins University Health Services Research and Development Center. Dr. Garnick is Associate Research Professor, Institute for Health Policy, Heller Graduate School, Brandeis University. Mr. Richards is Systems Manager, The Johns Hopkins University Health Services Research and Development Center. Dr. Fowles is Vice President, Park Nicollet Medical Foundation. Dr. Lawthers is Lecturer, Center for Quality of Care Research and Education, Harvard School of Public Health. Mr. Chandler is Programmer, The Johns Hopkins University Health Services Research and Development Center. Dr. Palmer is Director, Center for Quality of Care Research and Education, Harvard School of Public Health.

This work was completed as part of the DEMPAQ project. The DEMPAQ Research Team includes the authors of this paper as well as Jean Edwards and Duc Nguyen, Ph.D., of the Delmarva Foundation for Medical Care.

[1] Medicare beneficiaries include Americans over the age of 65, the disabled population and people receiving treatment for end stage renal disease (ESRD).

In 1990, the Health Care Financing Administration's Health Standards and Quality Bureau (HSQB) sponsored a multiphase study of the complementary use of claims and charts for measuring the quality of care provided by office-based primary care physicians to Medicare beneficiaries. In this study we summarize the key methodologies used in the DEMPAQ project[2] to develop a patient-population-oriented database derived from insurance transaction data. After providing a background discussion on the structure of the NCH, we outline our strategy to clean, improve, and construct integrated analytic records extracted from the NCH. Next, we introduce several claims-derived technical innovations to support current and future quality improvement initiatives. This article will be relevant not only to those who wish to apply Medicare data for quality improvement (QI) analyses, but also persons who are interested in the general application of private and public sector health insurance claims for the purpose of health service analyses.

CONTENTS OF THE NCH FILE

HCFA administers the largest single insurance claims database in the United States. Hospital admissions, physician office services, outpatient services, and nursing home and home health services are all captured in HCFA's databases. When begun in 1989 and unveiled in 1992, the NCH file project resulted in streamlined claims processing and the creation of a database that links all of the different types of claims together based on a common beneficiary identification (ID) number.

In this section, we focus on the structure and contents of the NCH. Our intent is to provide an overview of the file's architecture and features. Specific attention is given to the following subjects:

- Part A and Part B Medicare benefits versus institutional and noninstitutional claims
- Headers and trailers
- Advantages of the NCH over past Medicare databases
- An inventory of the NCH files
- The value of carrier-based provider files
- Beneficiary and provider IDs: The keys to developing an episode of care
- Identifying diagnosis and procedure information.

Part A and Part B Medicare Benefits versus Institutional and Noninstitutional Claims

Medicare defines its benefits package in terms of Part A benefits which include inpatient, hospice, and nursing home, and Part B benefits which include outpatient services (e.g., physicians and hospital outpatient departments). However, Medicare's claims data files are not structured by benefit package. Rather, HCFA pays claims to providers and organizes its data files into two separate categories: institutional and noninstitutional providers based upon who submits the claim, not the type of service performed. Institutional providers are facilities such as hospitals, nursing homes, hospices, or home health services. Noninstitutional providers include physicians, group practices, and freestanding laboratory services. Thus, for example, an eye examination given by a house staff physician in an outpatient department is an institutional claim, whereas the same eye examination provided in a physician's office would be a noninstitutional claim, although both are covered by Part B benefits.

Part B and Part A claims can be analyzed together for sets of patients or providers using beneficiary ID numbers or a unique provider ID numbers. Moreover, 100% of all claims are maintained on the NCH. As a result, a beneficiary who spends half the year in Phoenix and half the year in Minnesota has a complete record of all their reimbursed health services recorded at HCFA despite the two different sets of providers and data processors.

Understanding the path of a claim from provider to HCFA's main office provides insight into the contents of the NCH. The path of a Medicare claim starts when a provider submits a claim to state or region specific carriers.[3] If an institutional provider performs a Part A service, the claim is submitted to the carrier on a UB-92 form. A noninstitutional or institutional provider of a Part B service would submit a claim on the HCFA-1500 form. The carrier records all of the claims information from both forms and then forwards the claims record to one of six regional host sites in the United States. The host sites prepare the data for submission to HCFA's main office in Baltimore, MD. Along the way, key data fields are checked for completeness (e.g.,

[2] The study's name, DEMPAQ, represents its mission to "Develop and Evaluate Methods to Promote Ambulatory care Quality." For a general overview of the DEMPAQ project's objectives, methods, and philosophy, see Lawthers et al. (20). The DEMPAQ research team developed careful specifications to abstract from HCFA 2 years of claims for Medicare beneficiaries residing in the states of Alabama, Iowa, and Maryland. The DEMPAQ team consisted of multidisciplinary group of researchers from the Harvard School of Public Health, the Delmarva Foundation for Medical Care, The Johns Hopkins School of Public Health, The Park Nicollet Medical Foundation, and Brandeis University.

[3] For example, Blue Cross Blue Shield of Alabama is the carrier for all claims submitted for providers practicing in Alabama.

diagnosis and procedure codes). The net result is the NCH, a national database of all Medicare beneficiary health service transactions regardless of the location of a claim's submission.

Headers and Trailers

Each of the NCH's raw claims records has two components: a "header" and "trailers." The header portion of the claim record contains the beneficiary ID, claim type, dates of service, and Medicare contract information. The trailer portion is comprised of different data groups including:

- Diagnostic and surgical data group
- Procedure and financial data group
- Admissions data group
- Beneficiary data group.

Each data group may be repeated a variable number of times. For example, noninstitutional file records contain up to 99 financial data groups and up to 9 diagnostic and surgical groups. The trailers contain a variable number of data groups. As a result, claims records may have different lengths, and the location of the data group may be different for two claims of the same type. For example, the location of the procedure code on the claim record for an office visit with one diagnosis code will be different from the location of the procedure code on a claim record for an office visit with two diagnosis codes.

Advantages of the National Claims History Files over Past Medicare Databases

Five significant improvements distinguish the NCH from other Medicare claims databases:

- The most significant improvement is the availability of 100% of Part B claims for 100% of HCFA beneficiaries. Previously, HCFA's Part B Beneficiary Annual Data File, the so-called BMAD file used from 1985 to 1990, contained only a sample of 5% of all beneficiaries with 100% of their claims.
- Although earlier files were limited to the carrier servicing a resident's claims, now it is possible to describe the claims associated with a beneficiary no matter where the beneficiary resides (e.g., Vermont) and where they might seek service (e.g., Florida).
- The number of edit-checked data fields has been increased and the checks made more sophisticated. For example, ICD9 diagnosis codes fields are evaluated for completion as well as accuracy. Other edit-checked critical data fields for quality of care analyses include procedure code and Unique Physician Identification Number (UPIN).
- The lag period between date of service and date recorded on the NCH is significantly shorter than before. Approximately 99% of claims are available for analysis 3 months after the date of service.
- HCFA can array multiyear claims data with a common beneficiary identification number across time (6).

An Inventory of the NCH Files

Although the "NCH file" title implies only a single file, the NCH is actually a "family" of three databases linked by beneficiary ID, as indicated in Table 1. A description of the three follows with highlights of key issues relevant to using the data for quality of care analyses.

Institutional Claims Files. This file contains claims data primarily on hospital, nursing home, hospice, and home care center services. Outpatient procedures provided by a hospital outpatient department and ambulatory surgical center will be found on this database. Each of the NCH's raw claims records (institutional and noninstitutional) has two components: a header and a trailer. The header portion of the claim record contains the claim ID, beneficiary ID, principal diagnosis codes, and basic Medicare contract information. The trailer portion contains information on the procedures performed, provider ID,

Table 1

Components of the NCH Database

Component	Sample of Contents of File
Beneficiary file — HISKEW file	Beneficiary information: unique identification number, age, sex, beneficiaries state, county and zip code, marital status
Institutional claims file: *contains Part A (inpatient) as well as Part B (outpatient) services*	Claims submitted by institutional providers (e.g., hospitals, nursing homes, hospices): patient ID, provider ID, date of service, type of service, procedure and diagnosis
Noninstitutional claims file: *contains Part B (primarily physician) services*	Claims submitted by noninstitutional providers (e.g., physicians): patient ID, provider ID, date of service, type of service, procedure and diagnosis

the specific diagnosis code associated with the procedure billed, charges submitted, final reimbursement, type of service, and other information.

Noninstitutional Claims Files. This file contains data primarily on physician services provided to beneficiaries. Another important source of noninstitutional claims are free-standing laboratories. The claim ID, provider ID, beneficiary ID, and principal diagnoses are the key data elements found on the header portion of the claim. This information is taken from top part of the HCFA-1500 form. Trailer information, recorded on the lower portion of the HCFA-1500 form, includes data elements describing procedure, service date, diagnosis associated with the procedure, place of service, and cost. It is important to note that physicians provide services in many different settings: i.e., a physician's office, inpatient, outpatient, nursing homes, and a patient's home. As long as the physician (either solo or group affiliated) is billing for a procedure, it does not matter where the procedure actually occurred.[4]

Health Insurance Skeleton Eligibility Write-off (HISKEW) File. The HISKEW is the core database used to link Part A and Part B claims by beneficiary ID. This file contains information on a beneficiary's age, gender, residence (state, county, and zip code), marital status, and enrollment information. Socioeconomic status can be partially inferred by a participation in Supplemental Security Insurance (SSI) data field available on the file. Enrollment dates describe when a person started to use Medicare services with a specific beneficiary ID. Beneficiary terminations from the Medicare program, due to moving or death are recorded as well on the "inactive file"; a separate but identically structured "active" version of the HISKEW file describes current Medicare recipients.

It is important to note that a beneficiary may have more than one ID in a lifetime. For example, assume a husband and wife share the same base portion of their beneficiary ID; the husband's social security number (SSN). To distinguish between the husband and the wife, each has a Beneficiary Identification Code (BIC) describing their respective marital statuses. His BIC, A1, shows that he is the principal contract holder. Her BIC, B1, indicates that she is the spouse of the principal contract holder. Each person's BIC augments the principal contract holder's SSN; making each beneficiary uniquely identified. If after 10 years in Medicare, the husband dies, his wife's beneficiary ID would change in two major ways. The widowed wife's SSN would replace her husband's number as the base portion of her beneficiary ID. Also, her BIC would change to indicate that she is widowed. As a result, the widowed wife will have two separate beneficiary IDs and associated claims using both IDs. Strategies we used to define a unique nonchanging beneficiary ID are described in the section on creating analytic files.

The Value of Carrier-Based Provider Files

Physician characteristics can be linked to the analytic files by provider ID. Under contract to HCFA,[5] each of the Medicare carriers collect information on physicians' specialties, credentials, board certification, medical school, and practice location. From a health services research perspective, these variables can be used to determine if physician characteristics have any meaningful correlation with practice patterns. During the DEMPAQ project, we found these files essential for determining whether the provider ID recorded on a claim is associated with a physician's solo or group practice. Using the UPIN, the employer identification number (EIN), and information indicating whether or not a particular record was associated with a group or solo practice, we could distinguish a provider's group and solo practice IDs. It should be noted that without the use of carrier provider files, or substantial revisions to the NCH, group practices cannot be uniquely identified. An approach to identify groups practices is discussed in more detail later.

Beneficiary and Provider IDs: The Keys to Developing an Episode of Care

Developing episodes of care allows an analyst to evaluate the quality of medical services received in units relevant to the practice of medicine (1–4, 6, 7). After identifying a key event; the onset of disease or use of major medical procedure (1), all of the claims associated with the key event are then aggregated to develop a patient-specific episode summary. Linking claims records by beneficiary and provider ID is critical in episode development.

[4] For example, at one point we noticed the second largest state (after their home state) rendering treatment to Maryland beneficiaries was South Carolina. More intriguing, most all of the services occurred in nursing homes. After a careful analysis, we realized the claims were for laboratory services, provided for patients residing in Maryland nursing homes that used the services of a laboratory located in South Carolina.

[5] Although carriers supply the data constituting the NCH, the provider files are not a part of the NCH. We were given special access to the carrier-supplied provider files to complete the DEMPAQ project. Those wishing to repeat the methods we used to identify unique providers will have to request access to the appropriate files.

Although all claims can be linked to a beneficiary by a common ID, the current specification of physicians', facilities', and other providers' IDs pose problems to analysis because each provider can use more than one ID. For example, a physician can use one ID for his/her solo practice and a different ID for each group practice he/she participates in. Consequently, the identification of the claims associated with one provider ranges from burdensome to impossible. Our interim strategy used to remedy this problem as well as HCFA's long-term solution is discussed later as a technical innovation.

The following example provides a sense of the number and scope of raw claims transactions needed to construct an episode of illness record from institutional and noninstitutional files. Suppose there is interest in studying amputations associated with diabetes. First, we would pass through roughly several million claims records (per state, per year) using the institutional file to find all amputation admissions. The resulting amputation admission records would be matched with any patient associated with a diabetes diagnosis on the noninstitutional and institutional files to specify a study population. An episode record can then be constructed for each diabetic in the study population. The episode's focal point is the amputation admission. The episode begins with the most recent primary care physician consult, before the amputation, and ends with the consult and follow-up visits provided up to 3 months after the surgery. The outpatient consults and follow-up visits can be obtained from the noninstitutional file. Physician visits and consults to the patient, while in the hospital, can also be found on the noninstitutional file. Home care services received immediately after the amputation are located on the institutional file.

Diagnosis and Procedure Code Standards for the NCH

Diagnosis information is recorded using five-digit International Classification of Disease Codes, 9th Revision (ICD-9). Outpatient procedures are classified by the HCFA Common Procedure Coding system (HCPC—a modified version of the American Medical Association Current Procedural Terminology (CPT) System). Inpatient procedures can be described by Diagnosis Related Groupings (DRG) or HCPC. If a facility bills for an inpatient procedure, a DRG describes the treatment. If a physician bills for professional services accompanying an inpatient procedure, a HCPC describes the service provided. Any surgical inpatient procedure is also described by ICD-9 surgical procedure codes. In situations where a new procedure is being used and an associated HCPC code is not available, "local" HCPC codes are maintained by state or region-specific carriers. During our study period, a large portion of outpatient institutional records have no HCPC associated with the service provided. In our subsequent analysis, we recorded the charges and service use of these unnamed procedures, but were not able to determine their description. HCFA has corrected this problem as of April 1, 1991 (personal correspondence with HCFA staff).

TURNING RAW CLAIMS FILES INTO ANALYTIC FILES

The creation of the NCH has greatly improved the quality of Medicare data. Moreover, HCFA is constantly updating the quality of the NCH. However, despite these improvements raw claims data are rarely ready to use, "as is" for most analyses (4, 8–10). In this section, we outline a strategy for turning raw claims files into analytic files. Our goals were to remove extraneous or inaccurate information and prepare the claims for analysis.

Selecting the Population of Interest

The entire Medicare database contains far more patients than most analysts want to study. Because the DEMPAQ analysis focused only on beneficiaries who belonged to the age 65+ Medicare program, we discarded the claims of any patients with end stage renal disease (ESRD) and disabled persons under the age of 65 years. (Disabled persons over the age of 65 are automatically transferred to the standard Medicare contract at their 65th birthday.)

We further limited our study population to persons residing in Alabama, Iowa, and Maryland at any time during the year. The records of beneficiaries who died part way through the year were flagged and kept for later analysis. To determine whether a beneficiary was a continuous resident or died within the year of observation, we used two "snapshots" of the beneficiary file. The first snapshot of the beneficiary file was taken at the start of the study period, July 1, 1990. The second snapshot of the beneficiary file was taken at the conclusion of the study period June 30, 1991. Using the participation and termination dates from the two points in time, we could differentiate persons who were continuous Medicare recipients from beneficiaries who died within the study period or were ineligible at the beginning of the period.

When selecting beneficiaries who meet the criteria for inclusion in an analysis, it is necessary to take into account the fact that people can have changing IDs

during a lifetime. "BIC-equating" is a process by which the person's multiple IDs across time are recoded by matching the birth/gender/SSN to identify a unique beneficiary ID and track a patient over the course of one or several periods. To provide this feature, HCFA will change a person's beneficiary ID registered on a claim if it differs from the original ID at the start of a period. A cruder, yet more straightforward, process to remedy changing IDs is to match claim and beneficiary information based on a combination of birth date, gender, and SSN. These data fields are available on the HISKEW as well as the claims database. This is not a trivial issue, in a 1-year period, we found that approximately 16% of the total claims could be mismatched due to changing ID numbers.

Preparing the Claims Data for Analysis

The task of preparing the claims for analysis had three objectives: conversion of files into a form that allowed rapid extraction of fields and records; removal of adjudications or transactions claims not associated with service delivery; and selection of the study population. To achieve these objectives, we completed the following steps:

Convert Claims to a Fixed Format. Claims record trailers can have variable record lengths. (i.e., it is possible that each record may be of a different length). Variable length records are difficult to use for health services analysis. Therefore, they were converted to a fixed length. The header and trailer were combined to form a single record describing the service provided. This unit of analysis is commonly called line-item detail. In contrast to claim-level detail, line items reflect individual services provided on a given date. Whereas, several nonrelated items may be submitted together on a single claims form. A unique claim ID was assigned to enable re-linking line items submitted on the same claim.

Table 2 presents a state-specific summary of claims, line items, services, and costs associated with the study population. The differences between claims and line item detail described in the data cleaning steps is evident. A further disaggregation of utilization information is service-level information. Service information is contained in the line-item detail of a record as a variable indicating the frequency a specific procedure code listed on the record was performed. For example, a line item states a chest x-ray was performed on a given date of service. The service frequency field indicates how many unique chest x-rays were performed on a given day. It is important to note variations in this field can give rise to anomalies. For example, as seen in Table 2, Alabama and Maryland were found to have roughly twice the number of services in comparison to Iowa. Upon closer examination, we found the problem in Alabama and Maryland to be largely attributable to anesthesiology services.[6] Without identification of the anesthesiology anomaly, the number of services can be significantly overstated.

Table 2

Service Utilization and Cost Breakdown by State

Service Use and Cost Information per Beneficiary	State		
	Alabama $N = 415,723$	Iowa $N = 350,731$	Maryland $N = 389,765$
Total claims	6,484,340	4,795,853	6,238,481
Total line items	12,853,752	9,076,118	11,657,476
Total services	27,264,197	12,643,598	21,141,754
Total amount paid ($)	2.6 billion	1.4 billion	1.9 billion

Select Claims for the Study Population. Once the claims are "fixed," the beneficiary ID listed on the claim is matched to the beneficiaries selected from the HISKEW file. This process reduces the number of line items used to only those individuals included on our study. It is critical that BIC-equating, described in the HISKEW beneficiary selection step, is completed before this match is made. Beneficiaries selected from the HISKEW file with no claims history were not included in the sample population. As a result, the DEMPAQ project analyzed only heath service users.

Identify Adjudications. Claims databases are the by-product of the fee-for-service reimbursement system. One common feature of claims processing is the "adjudication" process. For example, take the following sequence of events:

1. A physician submits a claim and a record is created.
2. HCFA decides that the procedure could have been substituted by a less expensive treatment.
3. HCFA creates an additional claims record that is identical to the first, except that the payment is the negative amount of the original. This action "zeros" the amount paid.
4. HCFA creates a final record that is identical to the first, except payment reflects the cost of the less expensive treatment.

[6] The number of services field associated with anesthesia claims describes the number of minutes that an anesthetic was used. Normally, the number of services simply indicates the number of times the procedure was used in a given day. Most times, the number of services associated with a procedure is one (e.g., open bypass surgery). An early sign of the anesthesia problem is the presence of large number of services in units of 5 (e.g., 35 services translates into 35 min of anesthesia). If the amount of anesthesia given is important information, a quick fix is to identify all of the anesthesia HCPC codes and recode their services field to one.

The net result of this adjudication is three claims, representing one service performed. For the purpose of analysis, only one service should be recorded. In any analysis with claims data, this step should be a necessary precaution to avoid double or triple counting of claims. HCFA has since made available (as of April, 1994) only final adjudication records to researchers.

Out of Scope Claims. Once the claims were fixed and adjudication information was dropped from the final claims files, out-of-scope claims were deleted. For this study, skilled nursing facility (SNF), hospice, and home care institutional file claims were dropped from the analysis. Also, beneficiaries who received care for part of the year from HMOs were deleted from the analysis because claims dates are unavailable from HMOs.

CREATING ANALYTIC FILES

Logical units of service used in quality of care analyses (e.g., a visit, a person, or an episode) are usually made up of many transactions. Furthermore, for most analyses, data elements from other databases (e.g., beneficiary and provider files) must be combined with claims records. The goal of creating "analytic files" is to provide a database that can be readily used for research and management analyses. Using cleaned data records, we developed several analytic files described below.

Claims Extracts

The cleaned claims records were used to create "stripped-down" files suitable for use on a PC platform. The institutional file was split into two so-called "extracts": a Part A institutional file with line item information regarding the study population's hospital admissions, and a Part B institutional file with information on services delivered to the beneficiary in outpatient settings, but provided by an institution (e.g., a hospital outpatient department). In this way, many variables not germane to the analysis were eliminated. For example, the approximately 2,000-byte record of the noninstitutional NCH file was cut to just over 200 bytes. Most of the information eliminated was HCFA claims-processing information.

Developing An Enhanced Beneficiary File

The HISKEW served as the base for the development of person-specific analytic files. To this, additional information was added describing whether the beneficiary had died or was institutionalized three contiguous months within one year, their case-mix category, and each person's source of primary care.

Case-mix adjustment is a method used to account for differences in resource use associated with the health status individual patients. The Ambulatory Care Group (ACG) system, developed at Johns Hopkins was used as the primary method to case-mix adjust expected service utilization and cost described in physician specific profiles (11). The ACGs constitute a constellation of 51 unique patient categories that infer degrees of illness burden based on the number and type of comorbidities. A person is categorized into a single ACG code assigned based on the ICD-9 diagnosis codes submitted on a claim form by providers seeing patient. Usually, one year of encounters are used to assign an ACG. The number of comorbidities is also explicitly identified by Ambulatory Diagnostic Groups (ADGs), a midpoint assignment of the ACG case-mix assignment software. All of the beneficiaries in the study population were assigned ADGs based on their ICD-9 diagnosis information provided on the claims files.

The patient's primary care source (PCS) served as the main provider unit of analysis for our study. Each beneficiary's contacts were evaluated to find evidence of repeated visits to a single primary care provider. We developed an algorithm to identify a beneficiary's PCS, defined as the primary care physician (i.e., general practice, family practice, internal medicine, or osteopathic medicine) who had seen a particular beneficiary most frequently. This determination was based upon "face-to-face" office encounters in which the physician was likely to spend a significant portion of time with the patient. Face-to-face encounters were distinguished by a set of HCPC codes and the place of service reported on noninstitutional claims. Of the primary care physician face-to-face contacts associated with a beneficiary, the provider with the most contacts was designated as the PCS. Ties between competing providers (affecting less than 2% of all providers) were broken by charge information.[7] For this study, a PCS was further characterized as a primary care physician who was assigned at least 25 patients and had at least two visits from any of their primary patients during each quarter of the year (12).[8]

---

[7] The provider delivering more intense services was designated the PCS.

[8] Our decision to organize our analysis around the PCS was motivated by a desire to establish a medical practice denominator to which a patient is assigned. We also wanted to test the logistic plausibility of Medicare patients adopting a managed care "gatekeeper" approach by examining their existing patterns of care. Those wishing to use only the provider ID used on the claim should beware of problems identifying a unique physician or practice discussed in the text.

Table 3

Breakdown of Study Population

Study Population Denominator	State		
	Alabama	Iowa	Maryland
Total beneficiaries in state	474,332	384,625	447,145
Study population (health service users)	415,723 (100%)	350,731 (100%)	389,765 (100%)
Study population assigned a PCS	328,740 (79%)	277,727 (79%)	286,221 (73%)
Study population assigned a PCS with >25 patients and >2 visits a quarter	298,394 (72%)	254,914 (73%)	264,885 (68%)

Table 3 shows that slightly less than 80% of the study population (health service users) remains following application of the PCS assignment algorithm. Making a further restriction regarding the patient load of and continuity of PCS's practice further reduces the study population to just over 70%. The Maryland study population was reduced the most from PCS assignment. This could be attributable to several factors. Beneficiaries residing in Maryland, may seek more medical care outside of a physician's office (i.e., in emergency rooms or outpatient clinics) than Iowa and Alabama patients. Another reason may be that Maryland patients see more specialists than primary care providers than Iowa and Alabama beneficiaries. A sensitivity test expanding the definition of a PCS to include specialists increased the beneficiaries assigned a PCS in Maryland to about 80% of the study population. This may account for the difference between the states.

Table 4 provides a comparison of the case-mix of the Alabama[9] total beneficiary population assigned and not assigned a PCS. Case-mix strata include age categories, gender, and the number of comorbidities associated with each patient. The number of comorbidities is based on the number of unique diagnostic groups produced from the ACG case-mix classification system. The patients without an assigned PCS are younger and tend to have a greater proportion of males than the assigned PCS population. The pronounced difference between the two populations is the number of comorbidities. The non-PCS population is clearly healthier. Much of this result is attributable to the fact that there are roughly 50,000 nonusers (i.e., no claims history) in the non-PCS category. Although, one may argue that nonusers in Medicare may be actual self-pay patients, the percent of the Medicare population paying their own way has to be fairly small if beneficiaries are rational. Even discounting the effect of including nonusers, the non-PCS assigned population are still significantly healthier than those assigned a PCS. Inasmuch as the focus of DEMPAQ was methods to evaluate quality of care we feel the focus on the PCS assigned population is justified by their more intense experience with the health care system.

Accounting for physician-specific practice patterns motivated the incorporation of a patient-specific case-mix index. In Table 5, we display a randomly selected PCS's average cost per patient by several health service categories to demonstrate the policy implications of using unadjusted versus case-mix-adjusted average costs to rate the cost-effectiveness of a provider. Cost per patient estimates were based on "resource units." These resource units, described in more detail below as a technical innovation, were developed from the resource-based relative value scale (RBRVS) used by HCFA to price procedure codes. The resources described can be thought of as a measure of cost per patient (in terms of dollars). By using resource units, cost comparisons can be made while controlling for provider-specific or regional price differences. Three types of resource units are compared: (a) the provider's

Table 4

Case-Mix of Alabama Patients with and without PCS Assignment

Case-Mix Measures	Study Population		
	Patients with PCS (%)	Patients without PCS (%)	All patients (%)
Age Strata			
65-74 years	56	59	57
75-84 years	35	30	34
85+ years	9	11	9
Gender strata			
Male	36	43	38
Female	64	57	62
Comorbidity level[a]			
0	0	58	18
1	10	15	11
2-3	30	16	26
4-5	28	7	21
6-9	27	4	20
10+	5	0	4
No of beneficiaries	328,470	145,862	474,332

[a] Number of conditions. Based on disease clusters with ACG system.

[9] In the interest of minimizing the permutation of different States' information, the Alabama Medicare population was chosen as the focus of our results. Identical analyses can be presented for Iowa and Maryland.

own costs per patient based on the procedures billed for primary patients, (b) the state average cost per patient, and (c) the provider's expected costs per patient based on calculations using ACGs. For this provider, actual resources per patient are higher than the average. However, the ratio of actual to expected resource use is less than one. Taking the two comparisons together, the provider uses more resources than average but has a patient population that demands more intensive treatment.

The information displayed in Table 5 can also be expanded to determine if their are any substitutions between medical service resources. For example, if the actual to expected ratio for ambulatory services was 1.20 (indicating a resource use of 20% over average) and the actual to expected ratio for inpatient services was 0.80 one could conclude the provider prefers treating patients in an ambulatory setting. That is, the provider is substituting away inpatient service for ambulatory services. To determine if the physician is being more or less cost-effective some additional measure of health outcomes would have to be used to gauge whether the additional ambulatory resources spent at the margin produce a comparable health outcome to those that might have been used in an inpatient setting.

PCS Quality and Utilization Profile Files

Using the extract claims and enhanced beneficiary files, we created physician-specific profiles. These profiles represent provider-specific summaries of a patient's quality of care and utilization measures. Two types of profiles were generated for the DEMPAQ project: condition-specific profiles and office practice profiles. The condition-specific profiles described process and outcome measures associated with a PCS's patients for a given medical condition. The conditions analyzed included diabetes, COPD, congestive heart failure, osteoarthritis, hypertension and ischemic heart disease. The office practice profile focused on demographic, utilization and cost statistics associated with a primary care source's patients. Inpatient as well as ambulatory services were linked on a beneficiary level and aggregated to the provider level to create the profiles. All services—not just those used by the PCS—were included in the profile. Complete examples of the quality and utilization profiles are presented in Garnick et al. (13).

Table 5

A Sample Physician's Actual Resource Units per Patient Compared to the State Average and Case-Mix Adjusted Expected Resources

Resource Unit Category	Resource Use per Patient				
	Actual	State average	Actual to average ratio	Expected based on case-mix	Actual to expected ratio
Ambulatory services	1577.19	1493.23	1.06	1732.65	0.91
Laboratory services	102.26	77.09	1.33	88.26	1.16
Hospital services	2778.98	2419.29	1.15	2920.92	0.95
Total services	4458.43	3989.61	1.12	4741.83	0.94

A ratio greater than 1 indicates that the actual is greater than average or expected.

TECHNICAL INNOVATIONS

To create the physician profiles, several innovative strategies extend the abilities of the NCH database. These included the development of a resource unit table to uniformly price all services, the generation of solo and group practice identification numbers, and linking physician inpatient and hospital admission claims.

Resource Unit Table

In an attempt to minimize the bias of any region specific service pricing and reimbursement policies and to shield the physician's actual charges from being scrutinized, standard "resource units" were developed to reflect reimbursement for all of the types of procedures used in the claims data. Our resource unit calculation approach varied by different types of procedure billing codes:

- For *inpatient information*, standardized reimbursement was based on the average amount paid by HCFA per DRG in the state of beneficiary residence.[10]
- For *outpatient procedures*, the RBRVS weights for each HCPC, ignoring geographic factor price differentials, were multiplied by $31.00 (the 1992 RBRVS conversion factor) to produce standardized resource units (14).
- For *procedure codes listed in the claims but for which no RBRVS HCPC was available*, the state average reimbursement for the procedure was used.

[10] ICD-9 surgical procedures are also available on the inpatient claims. Unfortunately, the ICD-9 codes were inconsistent. Thus, DRGs were used exclusively to describe inpatient procedures billed by a facility.

- *Missing procedure* codes—which were concentrated in the institutional Part B claims—presented a problem. These records were associated with outpatient services provided in a facility rather than a provider's office. Although these claims were paid and have diagnosis information, they contained no consistent indication of what procedure was performed. For a small percentage of these missing procedures, a proxy measure is HCFA's revenue center code that describes, in the broadest terms (e.g., radiology) what procedure was performed. Because cost and utilization information from the claims data was necessary to create an accurate analysis and these nonprocedures line-items could not be priced (i.e., they were never associated with a fee), they were treated as one single procedure in the final summary of a primary care source's patients utilization and cost. We learned from HCFA that the non-procedure issue was confined to general medical services provided mainly by hospital outpatient departments (OPDs).[11] According to HCFA, as of April 1, 1991, all institutional Part B claims include HCPC codes.

Distinguishing Between Solo and Group Practices

In all three states, many physicians participated in more than one practice. An innovation of this study was the creation of identifiers to distinguish between the services a physician service provided on a solo basis, or as part of a group. The first step in identifying a unique practice was to assess whether the "Provider ID" field on the claim was unique to only a single physician or multiple providers practicing as part of a group. To test this, the provider IDs on the claims were matched to those in carrier-supplied provider files. As Table 6 shows, the number of unique provider IDs (13,229) used in Alabama noninstitutional claims was roughly two times larger than the actual number of physician UPINs (6,954) in the provider file. A set of recoded provider IDs was developed by matching provider IDs and looking on the carrier-supplied provider files for group practice identification. Identification consisted of matching all possible provider ID numbers to the carrier-supplied provider files for each state and then distinguishing group practices based on EIN and group practice indicator.[12] The EIN is the common tax identification number used by a group practice to report its income to the Internal Revenue Service. This method of recoding provider IDs provided a reliable method to distinguish unique medical practices for future profiling and analysis.

Table 6

A Comparison of Three Methods Identifying Unique Medical Practices Based on NCH Part B Claims and the Alabama Provider File

Unique Practice Identification Criteria	Provider ID Type for Alabama NCH Part B Claims		
	Submitted provider ID	Submitted UPIN	Recoded provider ID
Number of unique provider IDs	13,229	6,954	7,135
Number of solo practice physicians with multiple IDs	1,498	0	0
Percent of line items of missing IDs	0	34% of line items	0
No. of patient associated with missing IDs	0	346,511 (83% of patients)	0
Actual practices			
Solo	4,777	1,237	2,722
Group	5,388	4,843	2,614
Other/unknown	3,064	874	1,799

One alternative to developing our own provider ID, simply using the UPIN, was found to be a severely limiting option. Table 6 highlights the dangers of relying entirely on UPIN to identify a group practice. UPIN was a required field for only half of the time period associated with the study period. However, as seen in column 2, the UPIN's absence from at least one claims record associated with a beneficiary affected 83% of the study population. This extent of missing provider information motivated the decision to develop a different method to identify solo and group practices. With UPIN now a required field, some of this problem should be eliminated. However, without explicit identification of whether the claim is submitted by a group or solo practice, our method remains a burdensome alternative to correct group practice identification.[13]

Specialty information from the claims and the provider file can be further used to distinguish between

[11] It is important to note that the provider identified on an OPD claim for a physician service is the hospital, not the physician providing the service. As a result, assessing the practice patterns of physicians working for a hospital in an outpatient setting will not be possible without revision of the NCH database and probably the claim form used by the OPD.

[12] Iowa and Alabama beneficiary claims are processed by Iowa Blue Cross Blue Shield and Alabama Blue Cross Blue Shield, respectively. Maryland's claims are processed by two carriers: Blue Shield of Pennsylvania (Prince George and Montgomery county beneficiaries) and Blue Cross Blue Shield of Maryland (rest of Maryland). As a result, a provider file from each of Maryland's carriers had to be obtained to identify group practices.

[13] Without the use of the carrier-supplied provider file, practice type identification would have been impossible.

Table 7

PCS Practice Type Breakdown

Provider Type	State		
	Alabama	Iowa	Maryland
Total (PCS)	1,370 (100%)	1,194 (100%)	1,702 (100%)
Solo practitioner	815 (59%)	900 (75%)	1,440 (85%)
Group practice	555 (41%)	294 (75%)	262 (15%)
Primary care group practice	490 (36%)	202 (17%)	165 (10%)
Majority primary care group practice	65 (5%)	92 (8%)	97 (5%)

different types of group practices. As Table 7 shows, we distinguished three types of group practices. The first type is a group practice consisting of only primary care physicians. The next two types have a mix of primary care and specialty physicians. When more than 50% of the physicians are primary care providers, the group practice was denoted as majority primary care group practice PCS. (A practice below 50% primary care was considered a specialty practice and thus was not denoted as a primary care source.) It should be noted that the difference in number of physicians listed in Table 7 for Alabama is significantly smaller than in Table 6 due to a focus on primary care sources. Once the 25+ patients and greater than two visits a quarter rule is applied to the Alabama PCS providers, the number of practices is further cut from 1,370 to 866.

Linking Physician Inpatient Claims to Hospital Inpatient Claims

Many health services analyses require capturing the total inpatient costs per admission. Most claims systems, including those at HCFA, maintain hospital bills in one database and physician bills, including the physician component of an inpatient stay, in another. In order to obtain a total cost per admission, it is necessary to link the physician bills to the appropriate admission. This can be problematic due to the sizes of the claims databases and if, as is usually the case, the physician bills do not contain a reference, i.e., admission date, to the inpatient stay. An approach that was used to accomplish the linking of physician inpatient claims to corresponding hospital admission dates is described below.

The approach required that physician claims be matched to hospital claims to obtain pertinent admission information, e.g., admission date. An obstacle to this task was that the physician claims usually contain only a single date corresponding to the actual date that the physician provided service, whereas the hospital claims usually contain a beginning and ending date corresponding to the range of dates the patient was in the hospital. In this case, a match of single date to an implied range of dates needed to be accomplished.

The first step was to select record identification information from the physician claims: person ID, date of service and a record number. The second step in the process was to select pertinent information from the inpatient claims: patient ID, admission date and beginning and ending dates of service. Note that because most hospitals bill on a cycle and not by admission, the beginning date of service will not necessarily be the admission date, but rather the first day of the billing cycle.

The next step involved converting the selected inpatient records, one record per bill, to one record per day in the hospital. So, for example, if a record indicated October 1, 1993, as the beginning date of service and October 31, 1993, as the ending date of service, it would be converted to 31 separate records, each one containing the person ID, single date of service (which now corresponds to a single day in the hospital), and admission date.

Finally, we matched selected physician claims to the "days in the hospital" file created in the previous step. The admission date was attached to the physician claims. A resorted physician claims extract, containing the admission date, was then matched back to the original claims by record number. Following this process, we achieved a match rate of 93% for Medicare claims for Maryland in 1991, 87% for Alabama, and 91% for Iowa. The nonmatched physician claims are attributed to keypunch errors and missing inpatient claims (possibly due to adjudications).

ASSESSING THE ACCURACY OF NCH CLAIMS DATA

Our experience with HCFA's NCH file provides a preview of some methodological and technical hurdles that will be confronted when regional or national databases are derived from insurance claims. One major underlying concern throughout the DEMPAQ project was the accuracy of the data; particularly diagnostic

and procedural information. The quality of the diagnostic information supplied on physician claims seemed quite good. Two separate analyses highlight the relative quality of the ICD-9 codes used in the analysis.

The first analysis was a by-product of the case-mix adjustment process. The software that assigns ACG categories was unable to recognize less than 3% of the diagnosis codes found in the physician ambulatory claims. The ungroupable codes included blank, local codes used by carriers, or undecipherable codes. To contend with the possible problem of zero-filling the right-hand side of the diagnosis codes, an algorithm was devised to work backwards from five to four to three digit identification to group the diagnosis codes. The degree of zero filling did not produce appreciable differences in the case-mix distribution.

The DEMPAQ project also undertook a more formal test of the diagnostic accuracy of claims by comparing them to medical records. This assessment was based on a sample of 2,000 patients chosen at random from the Maryland study population. Chart information was abstracted for these 2,000 patients from their PCS's records and compared to the information within the computerized claims records associated with this population. Of the six profiled conditions, agreement on presence/absence of the six profiled conditions ranged from 74% to 96% with κ values from 0.27 to 0.67. In cases where the claims data indicated the condition existed, but the chart did not mention the condition, a second small review of charts found concordance with the claims data in 18 out of 21 cases sampled (15). In this latter review, the nurse was allowed to infer from the physician's notes whether the patient was under treatment for the condition in question, even though the diagnosis was not mentioned verbatim in the chart. For example, adjustment of insulin was considered evidence that a patient *did* have diabetes even though the diagnosis was not written in physician's progress notes. This comparison study indicated that NCH data are capable of achieving significant clinical concordance with chart data. Although the DEMPAQ research team only focussed on six medical conditions, those medical conditions constitute a third of all care for the Maryland study population [See Fowles et al. (15) for a further description of this reliability assessment.]

EXPECTED AND SUGGESTED ENHANCEMENTS TO THE NCH FILE

While working with the NCH file, we maintained active communication with HCFA regarding our findings and suggested improvements. The quality of the NCH file proved to be dramatically better than older Medicare files and, indeed, current NCH claims have been further enhanced beyond those used in this study. The latest generation of files contain: (a) fixed formats, (b) availability of "final action" claims reflecting only health services,[14] and (c) a significantly smaller size due to a removal of claims processing information data elements.

Several further enhancements were suggested to HCFA to further improve the quality of the NCH database:

- An edit-checked UPIN (against a national registry) on every claim.
- Collection of a group practice's EIN on the claim record as well as a flag for solo versus group practice practitioner.
- A unique nonchanging patient identifier (e.g., SSN)

An edit-checked UPIN provides a means to uniquely identify any physician despite their group practice affiliations. Physician profiling could focus on either the individual physician level or the practice level. At the moment, changing the denominator from physician to practice remains a fairly complex task with some possibility for error. The UPIN is based upon a national system of unique physician identification. Matching UPIN to a national registry will minimize the chances of aggregating information to the physician level inappropriately.

Collection of a group practice's EIN and a flag to distinguish a service performed by a provider wearing their group practice or solo practice "hat" would eliminate reliance on the carrier-supplied provider file for this level of information. Given the significant proportion of group practices acting as a patient's primary care source, it will be important to correctly identify their unique practice style. Making the carrier-supplied provider files available is an untenable option. None of the four files we encountered shared a common format. Moreover, given that one state's beneficiaries receive care from providers in many states, an analysis in one locale would require a provider database from multiple—if not all—states.

Future longitudinal analyses using HCFA data will be dependent upon correct identification of beneficiaries over time. At present, HCFA continues to use the combination of the principal contract number and BIC code to uniquely identify a beneficiary. If HCFA were simply to use each person's SSN as the primary component of a beneficiary ID, a person could then be identified across time without regard to their marital status and other factors.

---

[14] Administrative adjudication claims records have been eliminated in the "final action" version of the NCH. The final action file is available from HCFA by request.

TECHNICAL CONSIDERATIONS

The NCH is the largest claims database system in the world. Using this database requires clarity of purpose. That is, with data files this large, it is difficult to perform an exploratory data analysis, even with just a 5% sample. The DEMPAQ project emphasized three major task categories: (a) the generation of physician profiles, (b) "querying" the database to produce frequencies and simple statistics, and (c) the production of complex ad hoc data requests. Querying is defined as a simple question to be answered by using the database (e.g., what is the number of females with diabetes in a county). In contrast, ad hoc analysis is distinguished by addressing several questions or hypotheses. The technical requirements to use claims data for these three applications are discussed below.

Generating Physician Profiles

Development and Enhancement. To produce standardized physician profiles for an entire region access to a mainframe computer is necessary; either to select the data elements to download to a PC platform, or to process and analyze the files as a whole. (For DEMPAQ, profiles were generated using a combination of both approaches.)

Mainframe utility programs can be used to extract key data elements to be downloaded to a PC. Once on a PC, either programming/statistical software (e.g., SAS) or a relational database package (e.g., FOXPRO) can be used to generate the elements of a physician profile from the extracted claims, beneficiary and provider records. The complexity of generating a provider profile increases exponentially as case-mix adjustment and episode of care features are incorporated.

Printing. Once the elements of a physician profile are generated, the profile can be easily printed. Organizations wishing to develop a quick means of printing and sending a profile should first develop a common summary database that could easily be based on a PC platform. From this database, a user interface can be designed to simply specify the name of the provider and what components of a physician's total practice should be printed. Ideally, a CD-ROM technology would be well suited for this task. In the case of Medicare Peer Review Activities, HCFA could create the files internally or contract out the task of generating periodic physician profile summary files available for the PROs to use.

Querying the Database

Many times a "routine" profile is not sufficient to address a specific policy question. A query system of the NCH would be useful to generate specific, one-time summaries of several key fields in the database. For example, one could query how many angioplasty procedures were performed in a state in a given year. With a mainframe, several commercially prepared query systems are available. To develop a PC-based query system, one must have ample storage space (e.g., several gigabytes) and a fast processor. Our experience indicates that a parallel processing technique or some other innovation to dramatically increase processing speed is needed to make this a viable option. Based on the rate of innovation in information systems developments, a PC-based system would be a reasonable goal within a few years.

Complete Complex Ad Hoc Data Analyses

If a physician profiling or a simple query system are not sufficient resources to study an analysis with many hypotheses to examine, the remaining alternative is an ad hoc data request system. Given the size of the NCH files, (even on statewide level) an ad hoc approach can not be reasonably completed on anything less than a mainframe or a large minicomputer. This assumes an analyst will require the full set of claims to complete each analysis. For claims previously selected for a specialized population (e.g., coronary artery bypass graft patients), a PC-based approach for ad hoc analysis may be feasible, but still fairly limiting in terms of time resources.

POTENTIAL START-UP COSTS FOR NEW USERS OF CLAIMS DATA

Claims data are a valuable resource, yet the cost of using the data is not trivial. Potential users of claims data are continually faced with substantial capital and labor costs. Most organizations wishing to use claims data should set a goal to use claims-derived analytic files on a stand alone microcomputer or a PC network. To meet this goal, analytic claims files need to be created from raw claims. Raw claims data files are often large, complex, and usually require considerable preprocessing on a mainframe platform. For example, 1 year of the 5% sample of the NCH file contains over 40 million claims records with up to 58 trailers and variable lengths consisting of a maximum of 3,504 bytes

Preprocessing this file on a PC is possible, but it requires substantial information system expertise. Claims users can overcome these technical restraints without the purchase of a mainframe, but a considerable investment is required in compatible hardware and software.

Labor costs are an important consideration for researchers contemplating using claims data as capital expenditures. At present, there is no formal training program to instruct programmers how to utilize claims data for health policy research. Informal training is generally provided by pairing a capable programmer with a health services researcher. The start-up costs associated with this process are not trivial. Most programmers require several months to develop a sufficient breadth of knowledge of the database to complete an analysis. In addition, researchers without any experience in database handling will have to invest the time to learn a new vocabulary previously reserved to insurers, programmers, and computer scientists.

One of the unexpected findings of researchers using claims data is the marginal costs of using claims may not go down after the first project. Depending upon the nature of the second or later project, it is very possible that the marginal costs may increase. This finding is due to fact that there are few standards in claims data formats. Even for experts, each new database is a new puzzle. There is usually some systematic set of common data elements (e.g., provider and patient IDs, procedure, and diagnostic coding), but the subtle nuances of different types of transactions, adjudications, and coding always will remain and must be fully understood to produce an accurate analysis. Often a programmer must carefully scrutinize each new claims database to find the variables of interest and reconcile any problems with missing data and unwanted records. Yet, some of these time costs are likely to go down as a programmer and researcher define and continually update a "short-list" of the data elements required for an analysis that are generally available in most databases.

ENSURING CONFIDENTIALITY

In any study recording the medical care encounters of a population in a database, it is of paramount importance that the privacy and security of the data be maintained as a primary principle for research (16, 17). There are many institutional safeguards in place to ensure confidentiality of a patient's medical encounter such as institutional review boards and committees on human subjects. As the use of claims data is likely to increase as a critical component of a new and complex health care information infrastructure, adequate precautions must be taken. Although these data hold great promise to monitor health care utilization and quality, it is critical that nothing undermines the tenuous trust between the research and policy community handling and using the data and those in the population at large who have unknowingly become study subjects.

Given that confidentiality is essential, what methods are available to guarantee the privacy and security of claims data? The first step is to take any patient or provider identification and encrypt them into a form that can not be identified. Meux (18) provides a good approach to encrypt identifiers. Second, each researcher should develop a database privacy and security plan that includes the use of confidentiality statements and the delineation of key personnel responsible for securing the data in that minimizes the threat of damage or theft. The plan should also note how intermediary data files will be either destroyed at the conclusion of the project or stored in a secure area. In summary, privacy and confidentiality are likely to become key issues as more community databases are utilized and health insurers begin to contract with researchers to complete quality of care studies.

SUMMATION

The quality improvement process is an iterative one and dependent on information that documents change. Berwick (19) challenged researchers to develop easily applied methods and databases so that physicians, insurers and analysts could generate an informed interchange to improve health care delivery. Jencks and Wilensky (5) responded to Berwick's charge on behalf of HCFA, with the announcement of the Health Care Quality Improvement Initiative. The development of the NCH as a comprehensive database for quality improvement was one facet of this program strategy. However, the database is necessary, though not sufficient, for meaningful analysis. The subsequent preparation and application of these data to achieve the quality improvement goals at hand is a process with many decision points and little precedent. Such was the path-breaking intent of the DEMPAQ project.

As health reform unfolds, the demand for the information regarding the quality and efficiency of health services will increase dramatically. Private and public insurance agencies are aggressively seeking value for their health resource expenditures. To support this increased demand for information, comprehensive databases that focus on communities and patients and not on individual transactions must be developed. Health insurance claims data hold great potential for meeting many of these future demands. This experience with HCFA's NCH file illustrates the usefulness of such data

and highlights the need to treat them with special care. We hope that this description of the challenges encountered and methods used in this endeavor will aid others seeking to use administrative data for health care quality improvement initiatives.

ACKNOWLEDGMENTS

The authors especially thank Paul Elstein, Ph.D., and Peggy Bowen for their facilitation of essential technical and logistical support from HCFA. Special thanks to Nancy McCall, Ph.D., for her advice and consultations regarding the NCH. The views expressed are the author's and should not be associated with the Health Care Financing Administration or any participating organization.

References

1. Hornbrook MC, Hurtado AV, John RE. Health care episodes: definition, measurement and use. *Med Care Rev* 1985;42:163–218.
2. Anderson G, Steinberg EP, Whittle J, et al. Development of clinical and economic prognoses from medical claims data. *JAMA* 1990;253:967–972.
3. Fisher ES, Malenka DJ, Wennberg JE, et al. Technology assessment using insurance claims: example of prostatectomy. *Int J Technol Assess Health Care* 1990;6:194–202.
4. Garnick DW, Hendricks AM, Comstock CB. Measuring quality of care. *Int J Qual Health Care* 1994;6:163–177.
5. Jencks SF, Wilensky GR. The health care quality improvement initiative. *JAMA* 1992;268:900–903.
6. Warren JL, Babish JD, Nicholson G. Use and linking of medicare data bases in creating episode of care file. Presented at the American Public Health Association 1990 Annual Meeting.
7. Weiner JP, Powe NR, Steinwachs DM, et al. Applying insurance claims data to assess quality of care: a compilation of potential indicators. *Qual Rev Bull* 1990;16:424–438.
8. Lohr KN. Use of insurance claims data in measuring quality of care. *Int J Technol Assess Health Care* 1990:6:263–271
9. McDonald CJ, Hui SL. The analysis of humongous databases: problems and pitfalls. *Stat Med* 1991;10:511–518.
10. Newhouse JP. Medical effectiveness research data methods: summary and reactions. In: Grady ML, Schwartz HA, eds. Medical Effectiveness Research Data Methods. Rockville, MD: Agency for Health Care Policy Research, 1992.
11. Weiner JP, Starfield BH, Steinwachs DM, et al. Development of a population-oriented measure of ambulatory care case-mix. *Med Care* 1991;29:452–472.
12. Parente ST, Weiner JP, Fowles J, et al. Incorporating primary care source (PCS) assignment into medical practice variations analysis. *Med Decis Making*, October-December, 1992, Abstract.
13. Garnick DW, Fowles F, Lawthers A, et al. Profiling physicians' practice patterns: a new approach to assessing quality of care for the 1990s. *J Ambulatory Care Mgmt* 1994;17:44–75.
14. *Federal Register*, June 1991. Washington DC: US Government Printing Office.
15. Fowles JB, Lawthers AG, Weiner JP, et al. Agreement between physicians' office records and Medicare Part B claims data. *Health Care Finan Rev* 1995;16:189–199.
16. Gostin LO, Turek-Brezina J, Powers M, et al. Privacy and security of personal information in a new health care system. *JAMA* 1993;270:2487–2493.
17. Minard B. Health care computer systems for the 1990s: critical executive decisions. Ann Arbor, MI: Health Administration Press, 1991.
18. Meux E. Encrypting personal identifiers. *Health Serv Res* 1994;29:247–256.
19. Berwick DM. Continuous improvement as an ideal in health care. *N Engl J Med* 1989;320:53–56.
20. Lawthers AG, Palmer RH, Edwards JE, et al. Developing and evaluating performance measures for ambulatory care quality. *Jt Comm J Qual Improv* 1993;19:552–565.
21. U.S. Public Health Service. Health 1993. Washington, DC: U.S. Public Health Service, 1994.

THANK YOU

The *American Journal of Medical Quality* could not exist as a peer reviewed publication without the significant contribution of those who review the manuscripts we receive. In 1995, in addition to members of the Editorial Board, the following persons provided reviews for us and ultimately for our readers. To each of them we extend our sincerest thanks.

David L. Hawk, M.D., Robert E. Kramer, M.D., Herbert E. Segal, M.D., John P. Whiteley, M.D., and Michael Weitekemp, M.D.