SOCIETY OF ACTUARIES

RESEARCH PROJECT

GROUP MEDICAL INSURANCE CLAIMS

DATABASE COLLECTION AND ANALYSIS

REPORT FOR PUBLIC RELEASE

Prepared by

Kyle L. Grazier
University of Michigan

and

William G'Sell
Ann Arbor, MI

September 2004

SOCIETY OF ACTUARIES

MEDICAL LARGE CLAIMS EXPERIENCE COMMITTEE
2003 – 2004

Anthony J. Houghton, *Chairperson*
Dennis E. Daugherty
Charles S. Fuhrer
P. Anthony Hammond
Gordon Russel Hugh
John I. (Jim) Mange
Walter C. Marsh
Michael R. McLean
David E. Olsho
Brett A. Roush
Staff Liaison: John A. Luff

COMMITTEE ON HEALTH BENEFIT SYSTEMS RESEARCH
2003 – 2004

William R. Lane, *Chairperson*
Alan D. Ford
Charles S. Fuhrer
P. Anthony Hammond
Richard A. Kipp
Staff Liaison: Steven C. Siegel

ACKNOWLEDGMENTS

This study and report result from the efforts of many. We are grateful to the **Society of Actuaries** for its generous funding and ongoing support. We wish to thank the committee staff liaisons, **Jack Luff** and **Steve Siegel**, and other Society administrative and program personnel with whom we worked closely throughout the project: **Korrel Crawford, Kara Clark,** and **Julie Rogers**.

The members of the current and prior Medical Large Claims Experience Committees dedicated long hours to conceptualizing the study, recruiting participants, and reviewing analytic plans and reports: **Jim Ahrens, Dennis E. Daugherty, Alan D. Ford**, **Charles S. Fuhrer, P. Anthony Hammond, Gordon Russel Hugh, William R. Lane**, **John I. (Jim) Mange, Walter C. Marsh, Michael R. McLean, David E. Olsho, Susan Pierce, Brett A. Roush,** and **Leigh Wachenheim**. As Chairperson of the Committee during this and the prior study, **Anthony J. Houghton** willingly provided crucial input and offered readily available guidance throughout the study.

Of particular note are actuaries **Jim Mange** and **Brett Roush** who contributed much time to this study. They provided valuable insights into the study design, selection of analysis, validity of the findings, and content and format of the tables and databases. We are indebted to them for their quick and thoughtful responses to our questions and for their guidance.

A final note of thanks is due the anonymous data contributors. They invested significantly in the preparation and submission of the data, and in responding to queries throughout the study.

GROUP MEDICAL INSURANCE CLAIMS DATABASE COLLECTION AND ANALYSIS

EXECUTIVE SUMMARY

The Society of Actuaries (SoA) developed this research project to repeat and expand upon its "Group Medical Insurance Large Claims Database Collection and Analysis" study, which was published as a monograph in August of 1997. The general purpose of these studies is to discover and analyze factors affecting claim incidence rates and the distribution of claim sizes.

To oversee and direct the project, the SoA formed the "Medical Large Claims Experience Committee" (the advisory committee). This committee participated in preparing the data request and data specification that were sent to potential insurer participants. After the study was underway, through a series of document reviews and conference calls, the Committee guided the researcher in development of the database structure and the plan of analysis. In the latter stages of the project, as the researcher was conducting the planned analysis and developing presentations for this report, a Subcommittee (the review subcommittee), consisting of Jim Mange and Brett Roush, volunteered long hours to review and comment on the work in progress. The study was greatly enhanced by their hard work and dedication.

The prior large claims study analyzed claims from a group of insurers for two years: 1991 and 1992. The current study considers claims from three years: 1997, 1998 and 1999. While the former study considered only "large claims" (i.e., claimants with annual paid charges at least $25,000), the current study considers all claimants, regardless of charge amount. For the prior study, insurers submitted the claims of each claimant, aggregated to one record per claimant, by

totaling claim amounts and identifying one primary diagnosis for each claimant. For the present study, insurers submitted claims level data, which were aggregated by the researcher.

Eleven insurers contributed data in some form for the current study. The final submittal of data occurred in April 2002. The participants submitted a total of 86 source files containing about 16 gigabytes of data. One very small submittal was not in a form suitable for inclusion in the study, and it was omitted. The remaining ten submittals were processed to achieve their combination in a standard database format. During the course of analysis, data from three insurers were deemed unreliable, and their submittals were removed from the study. Data from seven insurers are summarized in the standardized, public release database, which occupies approximately 690 megabytes in its ASCII format. For the three years combined, the database contains 4,294,030 claimants and total paid charges of $7,068,612,616.49.

Most data sets were submitted on CD-ROM; although, zip drives and electronic file transfer were also used. Most files submitted were in ASCII code, either fixed record length or delimited files. Other file formats used for submittals include Microsoft Access Database, Paradox Database, and SAS Transport.

Compliance of submittals with the data specification varied from nearly complete to inadequate. Claims data were more available than exposure data. Nine of the eleven submittals included some exposure data; although, the form and completeness of submittals varied greatly. Ultimately, member exposure data from four of the seven insurers remaining in the database were deemed reliable enough for inclusion in the exposure analyses.

Many differences among submittals had to be reconciled in preparing the standardized database. There were also differences among claim files of a given insurer. The methodology section of this report describes many of the differences and their resolutions, and outlines the steps taken to clean and edit the data and to put it in standardized form. Appendix C describes the resulting public use database.

In close cooperation with the review subcommittee, the researcher refined and revised the plan of analysis as the work proceeded and the feasibility of the various proposed analyses were determined. The researcher and the review subcommittee worked through the planned analyses, exchanging ideas, results and draft tables as the product evolved. The results of these efforts are summarized in a ten series of tables provided as Microsoft Excel spreadsheets, which are incorporated by reference into this report.

Sections II through XI of this report reflect the titles of the various table series and describe the contents of each table. The spreadsheet files containing the tables are sequentially named "Table_II.xls" through "Table_XI.xls".

Few insurers were able to comply completely with the data specification. The detail of the specification made it difficult and time consuming for contributors to extract all requested information from their variety of information systems. The consequence of the incomplete and variable nature of the submittals was a very labor-intensive effort to interpret each submittal, to edit and clean the data, and to put the data in a standard form while retaining as much data as

possible. Despite these difficulties, a substantial database was constructed, containing approximately 4.3 million claimants and $7.1 billion in paid charges, for the three years combined. The database provided the basis for a set of useful analyses, presented as the tables in this report.

If there is a desire to conduct similar data collection and analysis in the future, the researcher recommends seeking out insurer participants committed to an ongoing, repeatable effort. The objective would be to establish methods, which would allow data collection and analysis to be quickly and efficiently repeated for future claim periods. The ongoing nature of the project, with data submitted periodically by the insurer participants, would permit development of consistent methods for standardizing, analyzing and reporting the data, compared to the relatively ad hoc approaches of this study and the prior study. The result should be decreased data loss and more rapid database construction and reporting.

To implement this approach, data submittals of the various insurer participants need not be identical. The researcher could develop standard methods for manipulating the data of each insurer into a common form. This effort would be worthwhile because of the project's ongoing nature. Once methods were established for a particular insurer, processing of future submittals would be much more efficient. The resulting database, which would grow over time, could develop into a very valuable resource for research.

TABLE OF CONTENTS

I. HISTORY, METHODOLOGY AND FINDINGS

A. Background and Overview

In October 1998, the Society of Actuaries (SoA) began to develop this research project, to repeat and expand upon its "Group Medical Insurance Large Claims Database Collection and Analysis" study, which was published as a monograph in August of 1997. The general purpose of these studies is to discover and analyze factors affecting claim incidence rates and the distribution of claim sizes.

The prior study was initiated in August 1992 and, for a group of insurers, analyzed claims from two years: 1991 and 1992. The current study considers claims from three years: 1997, 1998 and 1999. While the former study considered only "large claims" (i.e., claimants with annual paid charges at least $25,000), the current study considers all claimants, regardless of charge amount. For the prior study, insurers were requested to submit the claims of each claimant, aggregated to one record per claimant, by totaling claim amounts and identifying one primary diagnosis for each claimant. For the present study, insurers submitted claims level data, which were aggregated by the researcher.

From approximately August 1999 to January 2000 SoA staff and members identified potential data contributors for the current study. From November 1999 through May 2000, the SoA sent potential contributors a formal request to participate in the study, consisting of a cover letter, a list of data specification issues, and a claims study questionnaire (see Appendix A).

During development of the project, the researcher informally participated in discussions and review of documents. In February 2001 the SoA entered into a written contract with the researcher to conduct the study. In an effort to maximize the number of insurer participants, an initial deadline for data submission was extended as the project proceeded. The last submittal of data occurred in April 2002. In the meantime, the researcher proceeded with preliminary data processing and review as submittals were received. Eventually, eleven insurers contributed data in some form.

To oversee and direct the project, the SoA formed the "Medical Large Claims Experience Committee" (the advisory committee). Through a series of document reviews and conference calls, the Committee guided development of the database structure and the plan of analysis. In the latter stages of the project, as the researcher was conducting the planned analysis and developing presentations for this report, a Subcommittee (the review subcommittee), consisting of Jim Mange and Brett Roush, volunteered long hours to review and comment on the work in progress. The study was greatly enhanced by their hard work and dedication. This report and its associated tables, along with the accompanying database, represent the culmination of these collective efforts.

B. Methodology

1. Data Collection

The SoA prepared a detailed, comprehensive data specification to accompany its requests for insurer participation (See Appendix B). Three general categories of data were requested: block of business, exposure and claims. A "block name" was to be included in the block definition and in each exposure record, permitting an insured to be associated with a block of business. Identification numbers for the primary unit and the primary insured were to be included in each exposure and claim record, permitting a claimant to be associated with her exposure record and block of business.

No two data submittals are alike. Data submittals' and compliance with the data specification varied from nearly complete to inadequate. This report highlights some of the notable similarities and differences.

Most data sets were submitted on CD-ROM, but zip drives and electronic transfer (e-mail attachments and ftp) were also used. Most submittals were in ASCII code (American Standard Code for Information Interchange), either fixed length format or delimited (comma or tab) format. Other file formats used for submittals include Microsoft Access Database, Paradox Database, and SAS Transport.

The eleven insurers submitted a total of 86 source files containing about 16 gigabytes of data. The 51 claim files contained 13 gigabytes of data, and an additional 35 exposure and block files contained about 3 gigabytes, in the source file formats. One insurer, submitting a claim file of negligible size and in a form not capable of inclusion in the study was initially omitted. The data of the remaining ten insurers were processed and standardized. In the course of analysis, as explained later in this report, data from three of the remaining insurers were discovered to be questionable and was then removed from the database and report.

Obviously, many of the participating insurers submitted multiple claim and exposure files. One insurer, which probably came closest to matching the data specification, submitted 32 files: 29 ASCII files, one Microsoft Access file and two Microsoft Excel spreadsheet files. Not only did file formats differ among insurers, formats often differed among the various files of a single insurer.

Completeness of requested information was variable. A subscriber identifying number was always provided. Gender and relation were usually provided, as were birth dates. However, there was an insurer, later removed from the study, which provided gender for some of its claim files, but not for others. Another insurer, remaining in the study, did not provide gender in its claims files. That insurer did provide some gender information in its exposure data, but its submittal did not permit matching of its claim and exposure files.

One insurer, later removed from the study, submitted birth dates with some claim files and not with others. Another insurer provided age, rather than birth date. Dates were submitted in many

formats, differing among insurers, often differing among files of the same insurer, and sometimes differing within a given file. For example, dates were submitted with and without leading zeroes on month and day numbers. Dates were submitted with and without separators between year and month or month and day. Both slashes and dashes were used as separators.

Charges in the claim submittals sometimes included commas in the numbers. One insurer led each charge value with "$". Some insurers expressed charges in dollars, using a decimal point followed by two figures. Others used an implied decimal, presenting charges in cents. Some insurers submitted hospital, physician and other charges as separate records. Other insurers had fields for each charge type in the same record.

Diagnosis was provided more consistently than was procedure. Diagnosis and procedure code lengths varied among insurers. Diagnosis codes sometimes included "." after the third position, and sometimes they did not.

With the exceptions of the insurer eliminated from the study before data processing began, and one small submittal later removed from the study for other reasons (as explained later in this report), all participants submitted some kind of exposure information. The form and format of the exposure submittals varied greatly. One insurer provided a list of members with some individual characteristics, but with no associated time period or dates of coverage. This insurer was ultimately dropped from the study for other reasons, explained below. Two insurers provided some exposure data for subscribers, rather than for all members. One of these insurers was dropped from the study, as explained below. Six insurers provided some exposure data for

members. One of these insurers did not provide relation (to subscriber) for its claimants. Submittals for an additional two of these insurers showed more claimants than members. These two insurers were retained in the claims database, but their exposure data were not deemed credible, and they were omitted from the exposure analysis.

Submittals of exposure data varied greatly in approach. One insurer provided a record for each member for each month and year of coverage. Each record was identical except in values for coverage month and year. A continuously covered member could be expected to have 36 records. One member was represented by 72 records, pairs of which also differed by group identifier. One insurer provided exposure as the sum of member-months of coverage, by gender and age range, for each month of the study. Several insurers provided a record for each member, listing coverage effective and termination dates, consistent with the data specification. Some insurers had multiple records for an individual, for different periods of coverage.

Requested details of plan coverage were provided fairly completely for one insurer. Several provided partial information, while others provided none. Smoking status was provided by only one insurer, which was able to provide status for 25% of its member exposure file.

The ability to match claimants to members was variable. Two submittals resulted in match rates of 90% or more. Others had a match rate of zero; subscriber identifiers in claim and member files were formatted inconsistently.

2. Data Review and Editing

Data were reviewed for unexpected or unreasonable values, and for inconsistencies. In a very few cases, unreasonable records were removed. More importantly, four of eleven submittals were ultimately removed from the study. Two of these submittals were very small, but the other two were not. Some insurers submitted some claims occurring outside the study period; naturally, such claims were not included.

Dates were submitted in a variety of formats. Some submittals required creating multiple date masks and date fields, which were later consolidated, to capture dates of different format within the same file. A few unreasonable birth dates were eliminated. A calculated age of 108 is the maximum left in the data. Age is calculated as of the beginning of a claim year, by subtracting birth year from claim year.

One insurer submitted age, rather than birth date, in its claim files. Many apparent individuals were listed in the claim files with two ages in the same claim year, usually one year apart. Age appeared to be calculated as of each claim date, so that a person filing claims before and after her birth date would appear in the file with two different ages. For this case, birth year was computed as the claim year minus the larger of two ages which differed by one year A few claimants with three values for birth year were manually edited. If the three years were one year apart, the middle age was used, otherwise, the age values were blanked.

Identifying the claims associated with an individual claimant can present challenges. Four submittals included means to identify individuals without relying on a unique combination of subscriber identifier and relation (for subscriber or spouse), combined with gender and birth date for dependents. One of these submittals included unique member numbers. Two submittals provided a separate field which, coupled with the subscriber identifier, indicated an individual. One submittal contained sequentially numbered values in the relation field for dependents associated with a given subscriber identifier, allowing individual dependents to be distinguished. For other insurers, an individual claimant is taken to be associated with each unique combination of subscriber identifier, relation, gender and birth date. Since only one subscriber or spouse should generally be associated with each subscriber identifier, gender and birth date should not be required in order to distinguish subscriber or spouse claimants. One flaw with this approach is that the claims of same sex twins would be aggregated as one claimant.

So that the claims of one individual will not be aggregated as if they were for several claimants, it is important to review the data for inconsistencies and to eliminate them. To this end inconsistent values for gender or birth date are blanked for subscribers and spouses; they are blanked for all relations for the four submittals permitting separate identification of dependents. For submittals not explicitly distinguishing dependents, this task cannot be accomplished for dependents; dependents of different gender or birth date share a subscriber identifier. One insurer did not include gender in its claims files. Most of its dependents would be properly identified by birth date; however, for this insurer opposite sex dependent twins, not just same sex twins, could be aggregated as an individual.

In addition to blanking inconsistent values, missing values were filled in when possible. When gender or birth date was blank on some claims but not on others, for a claimant who could be identified by other means, such as subscriber identifier combined with a member identifier or relation (for subscriber or spouse), and for whom the available data were consistent, the known values were carried over to fill in the missing values. In addition, blank values of relation were filled in for two insurers who provided member identifiers.

Different submittals used different values to represent characteristics such as relationship and gender. A system of consistent values was developed, and values in the claim files were changed, as necessary, to comply with it. Claimants identified as dependents are primarily children. A user of the database may note that a few non-standard diagnosis codes are present. Such codes were kept in the claim files, and they appear as values in the database when they were one of the three highest cost codes for an individual. In analysis and in determining diagnosis category, such values were treated as missing values. In communication with the advisory committee, the length to retain in code fields was determined. Diagnosis, Procedure and Zip codes were each truncated to three characters.

Exposure files required a process similar to that described above for interpreting and cleaning the data. Inconsistent values were blanked for fields such as group identifier, type of coverage, type of contract, customer, sic, zip code, birth year, month or day, and out-of-pocket limits. The objective was to associate each member with a unique combination of these characteristics.

A fair amount of interpretation was required in distilling the exposure data. Submittals varied greatly. Multiple records for a member were common, and often were part of the data scheme. Inconsistencies in coverage effective and termination dates were not uncommon. One insurer provided exposure summaries at January 1997, at May 1998, and at January and December of 1999. The summaries for 1997 and 1998 were used for those claim years, respectively. For 1999 the researcher took data for members appearing only in the beginning 1999 file and for members only in the end 1999 file. In addition, the data from the end 1999 file were used for members present in both files. The assumption was that all members in either file should be considered covered in 1999 for the periods indicated by effective and termination dates in their records. It was further assumed that data from the end 1999 file were updates to the earlier file for claimants appearing in both files but having different values in some field(s).

3. Standardization and Aggregation

Claim files from each insurer had to be expressed in a common structure and format to allow their assembly into a common database for analysis and submittal. The approach was to define a standard record with fields of type and length adequate to accommodate the data field for any insurer (e.g., Subscriber ID lengths varied among insurers, so the longest length was used for the master database). Depending on the form of the data for each insurer, sub-files of claims data were created, containing the data values required for the master database. These sub-files were then matched, by claimant, on the master database, and the relevant data were copied from the sub-files to the master database.

The master database was initially developed by extracting and appending to one another the individual characteristics of claimants from each insurer's submittal (i.e., subscriber identifier, member identifier (if provided), relationship, gender and birth date). To these fields were added an insurer identifier and the claim year (so the claims of different insurers and years could be assembled into a common database). Blank fields were defined for the charges, diagnoses and procedures to be obtained from the claim sub-files. Additional blank fields were defined for fields anticipated for analysis and expected to be obtained from the exposure data (e.g., type of managed care, SIC, underwriting, out-of-pocket limits). Claim and exposure sub-files were developed for each insurer, and they were matched on the master database by claimant characteristics to fill in the blank fields.

For example, charges in an insurer's claim files were subtotaled by claimant characteristics, and the results were transferred to the master database as total paid, allowed and covered charges for hospital, physician and other services, to the extent these charge categories were provided by the various insurers. Some insurers provided hospital, physician and other charges in one record. Other insurers provided only one type of service charge per record, providing a code field indicating the service type. Most insurers involved the addition of hospital, physician and other charges to obtain the total charges. Methods of arriving at the appropriate subtotals, and the procedure for developing sub-files and matching them to the master database, depended on the nature of a submittal.

Unlike the prior large medical claims study, the current study identified primary diagnoses during aggregation of the data to one record per claimant per year. For the prior study, insurers submitted data that had already been aggregated to one record per claimant, and they determined primary diagnosis as part of the preparation of their submittals.

The current study was designed to identify three principal diagnosis codes for each claimant. In addition, it identified three principal procedure codes for each claimant. It also identified a primary diagnosis category (a collection of diagnosis codes) for each claimant. While developing the plan of analysis, some advisory committee members expressed an interest in exploring procedures as a function of diagnosis. Therefore, primary and secondary procedure codes were identified for the primary diagnosis category of each claimant. For each of these nine codes or categories (3 diagnosis codes, 1 diagnosis category, 3 procedure codes, 2

procedure codes for the diagnosis category) subtotal charges for claims with each code or category were also computed and placed in the database.

The typical method for determining principal diagnosis codes and primary diagnosis category are described here. Similar methods were employed for procedure codes. These methods were separately applied to each insurer's files.

For each insurer, paid charges were subtotaled for the group of fields identifying individual claimants and for each diagnosis code present in any claim of that individual. To these fields were added diagnosis category and its charge, and fields to contain the rank of diagnosis code and category, all with blank values. The diagnosis category description field was filled in by matching the diagnosis code with the diagnosis code range in a diagnosis category definition file. The diagnosis category charge was filled in by copying the charge for the diagnosis code (these amounts were later subtotaled by diagnosis category). This file was then sorted, first on individual characteristics, then on charge subtotal for the diagnosis code, then on the ICD-9 code (if consecutive records for an individual have the same charge amount, the alphanumeric sort order of the ICD-9 value determines the rank). The diagnosis code rank field was then filled-in, in accordance with the sort order. The claim sub-file was matched to the master database by individual characteristics, insurer and claim year, and the codes and subtotal charges for the three highest cost codes were transferred to the master database.

This claim sub-file was then reduced to eliminate diagnosis code, diagnosis charge and diagnosis rank fields and to subtotal diagnosis category charges by individual and category. The new file

was sorted on individual characteristics, then on diagnosis category charge and diagnosis category description, and the diagnosis category rank field was filled-in. The new claim sub-file was matched on the master database by individual characteristics, insurer and claim year, and the diagnosis category descriptions and subtotal charges for the category were transferred to the master database.

Note that a primary diagnosis code may be blank but have associated charges, if the charges for all blank codes add up to a sum greater than the sums for other codes. For the same reason, secondary and tertiary codes need not be blank for such a claimant record. Similarly, primary diagnosis code may be blank, with primary diagnosis category not blank. This situation results when a blank code field has charges higher than any specific ICD-9 code, but when several specific ICD-9 codes within a diagnosis category have collective charges greater than those for the blank code combined with anomalous codes defined within the "unknown" diagnosis category.

Similar methods were employed to include in the database the three highest cost procedure codes and their associated charges. In the course of analysis, these fields were dropped from the database because they were not providing particularly useful information.

Because some members of the advisory committee expressed an interest in exploring the possible relationship between procedure and diagnosis, the researcher determined the primary and secondary procedure codes, and associated charges, for each individual's primary diagnosis category. Each individual's primary diagnosis category was matched on the diagnosis category

definition file to determine the ICD-9 range. The ICD-9 range and the individual's characteristics were further matched on the insurer's original claim file to obtain subtotal charges for each procedure within the primary diagnosis category. The resulting file was sorted on individual characteristics and on the subtotal procedure charges. The rank was determined from the sort order, and the two highest cost procedure codes and charge subtotals for the primary diagnosis category were transferred to the master database. In the course of analysis, these fields were dropped from the database. The data were not complete enough to be useful.

When it was possible to link claim files with exposure files, values from exposure fields anticipated to be used for analysis were pulled into the master claims database. Such fields included: type of managed care, SIC, Underwriting Class, Out-of-Pocket Limit.

Flag fields were added to the database to indicate whether a claimant was covered by a PPO and to indicate whether a claimant is from an insurer the data of which is included in member exposure analysis.

To protect identities and confidentiality, a unique, sequential numeric identifier was created for each claimant in the master database. Insurer identifiers, subscriber identifiers and member identifiers were then removed. Claim year was retained so that annual claim files could be combined into a common database. Claimant identifiers were first assigned for claim year 1997, by descending order of a claimant's total paid charges. This approach scrambled insurers, so that their claims would not be blocked together in the database. For 1998 claimants with claims in the 1997 file, the claimant identifier was carried over from the prior year. Remaining claimants

were sorted by total paid charges and new identifiers, beginning where the prior year ended, were assigned in order of descending total paid charges. Claimants in 1999 who had prior year claims had their identifiers carried over, and new claimants again had identifiers assigned in order of descending paid charges.

C. Analysis and Results

In a series of conference calls, the researcher and the advisory committee developed a plan of analysis, describing the analyses and presentations to be attempted as part of this study. The participants recognized that results feasible for inclusion in the report would depend on the extent to which data submittals complied with the data specification. A subcommittee of the advisory committee was established to review analysis and tables as the work proceeded and to guide the researcher in revising the plan of analysis.

As anticipated, some variables were too incomplete to permit some of the proposed analysis. To protect the identity of individual insurers, the researcher and the review subcommittee established a requirement that data from at least three insurers would be required for each insurer subset that might be used for comparative analysis. For example, the only distinct "type of managed care" provided by at least three insurers was the PPO plan type. Therefore, analyses by plan type distinguished only PPO from all other plan types combined. Claimants grouped with non-PPO plan types include those covered by HMO, Point of Service, or Indemnity plans, as well as claimants for whom plan type was not identified.

Results of analysis are presented as ten series of tables. Each of these table series is presented as a separate Microsoft Excel spreadsheet file. These files are incorporated by reference into this report. The spreadsheet files are sequentially named "Table_II.xls" through "Table_XI.xls". Section headings in this report reflect the table names. The content of each table is described in the corresponding section of this report.

The tables are provided as spreadsheets to enable a user to extend an analysis, using summary data that have already been developed. To prevent inadvertent changes, the spreadsheets are protected and are "read-only". A user wishing to make changes to a worksheet should first remove protection using the "unprotect sheet" command on the "tools" menu.

These tables contain as many as 33 pages each (Table VII). Each table is designed to be paginated separately. For a table containing multiple worksheets, printing "entire workbook" should result in the correct pagination of the table.

Data submittals of three insurers were discovered to be questionable during the course of analysis, and these insurers were dropped from the database. In communication with the review subcommittee, the researchers conducted in-depth reviews of these submittals, and the consensus was that their data were unreliable and risked skewing the results of analysis.

Several data fields prepared and included in the database used for analysis were dropped from the public use database because their content was not complete or reliable enough to be useful. Group identifier was dropped because its format varied among insurers in ways that might permit their identification, and because it was not used in analysis.

More significantly, ten fields for procedure codes and their associated charges were dropped because the provision of meaningful codes was too incomplete. The dropped fields were the primary, secondary and tertiary procedure codes for each claimant, and the primary and

secondary procedure codes for each claimant's primary diagnosis category, along with the associated charge subtotals for each of these codes. Primary procedure code values were blank, "XXX" or "000" for 46% of claimants and for 69% of primary procedure charge totals. The vast majority of primary codes for high cost claimants had one of these values.

D. Recommendations

The Society of Actuaries and the advisory committee prepared a comprehensive data specification intended to link medical claims to membership/exposure information, to plan characteristics and to block of business. Few insurers were able to prepare submittals closely complying with the specification. A similar observation could be made for the prior large medical claims study, the ambitions of which were lower than were those for the current study.

The consequence of the incomplete and variable nature of the submittals is a very labor-intensive effort to interpret each submittal, to edit and clean the data, and to put the data in a standard form which retains as much data as possible. Despite the imperfect nature of the submittals, a substantial database was constructed, containing approximately 4.3 million claimants and $7.1 billion in paid charges, for the three years combined. The database provided the basis for a set of useful analyses, presented as the tables in this report.

If there is a desire to conduct similar data collection and analysis in the future, the researcher recommends seeking out an initial group of insurer participants committed to an ongoing, repeatable effort. The objective would be to establish methods, which would allow data collection and analysis to be efficiently repeated for future claim periods. Increased communication and feedback with the insurers could help assure that initial submittals are adequate. Subsequent submittals could be refined as a result of enhanced, ongoing communication.

The project could begin with a relatively small number of insurers willing to work closely with the researchers in developing feasible data specifications. The effort could be to define a data set which could be provided by the initial insurers, and which would be expected to be generally available for possible future participants. As the project proceeded, other interested insurers could be invited to join the effort. The ongoing nature of the project, with data submitted periodically by the insurer participants, would permit development of consistent methods for standardizing, analyzing and reporting the data, compared to the relatively ad hoc approach of this study and the prior study. The result should be decreased data loss and more rapid database construction and reporting. The accumulation of data over time, and, hopefully, with the addition of insurer participants, could prove very valuable.

To implement this approach, data submittals of the various insurer participants need not be identical. Such an expectation may not be practical, given the variety of data systems used by the participants. Instead, the researcher could develop standard methods for manipulating the data of each insurer into a common form. This effort would be worthwhile because of the ongoing nature of the project. Once methods were established for a particular insurer, processing of future submittals would be much more efficient.

Significant effort by insurer participants may be required, especially during a participant's initial involvement in the project, and one might consider the incentives for participation. It is expected that one product of the project would be a regularly updated, public use database, devoid of references to individual insurer participants. As an incentive, participating insurers could be offered additional, more detailed analysis of their own data, compared to the other collective

insurers. Moreover, a participant could be permitted to request custom analysis of its data, compared to the entire dataset. Hopefully, availability of such analysis would have enough value to attract additional insurer participants over the course of time.

II. ANALYSIS BY DEDUCTIBLE AMOUNT: CLAIMANTS, PAID CHARGES EXCEEDING DEDUCTIBLE, AVERAGE EXCESS CHARGES

This series of seven Tables (each presented as a separate worksheet) summarizes claims for nineteen deductible amounts ranging from $0 to $500,000. A table presents a separate block of data for each of the three study years. For each deductible amount, a table presents the number of claimants having a charge total exceeding the deductible, the amount of those claimants' charges exceeding the deductible (i.e., the deductible is not included in this amount), the percentage of total charges represented by the excess charges, and the average excess charges per claimant having charges exceeding the deductible. In addition, the rate at which claimants exceed each deductible amount is expressed as claimants per 1,000 claimants, per 1,000 members or per 1,000 subscribers, depending on the table under consideration.

Table II-A

Table II-A summarizes claim data for all claims of all insurers, regardless of whether exposure data were provided. The rate at which charges exceed the deductible amount is expressed as claimants per 1,000 claimants.

Table II-B

Data presented in Table II-B are limited to claims from a subset of four insurers providing credible exposure data for members. Because one of these insurers did not provide claimant relation to subscriber, these data are not broken-down by relation. The rate at which charges exceed the deductible amount is expressed as claimants per 1,000 members. In addition to

information presented for all tables, this table adds a column summarizing the average excess charges per covered member.

Tables II-C-1 through II-C-5

Data presented in Tables II-C are limited to claims from a subset of four insurers providing credible exposure data for subscribers. Three insurers from this subset are also among the four insurers providing credible member exposure data. Each of the insurers providing subscriber exposure also provides claimant relation to subscriber, permitting the data to be broken-down by relation. Tables II-C-1 through II-C-5, respectively, summarize data for all members, for subscribers only, for spouses only, for dependents only (primarily children), and for unknown relation. The rate at which charges exceed the deductible amount is expressed as claimants per 1,000 subscribers. In addition to information presented for all tables, this table adds a column summarizing the average excess charges per covered subscriber.

III. ANALYSIS BY DIAGNOSIS CATEGORY AND DEDUCTIBLE AMOUNT: CLAIMANTS, PAID CHARGES EXCEEDING DEDUCTIBLE, AVERAGE EXCESS CHARGES

Table III summarizes, by diagnosis category, paid charges exceeding deductible for the claims of all insurers combined. Data are presented on separate pages for each of six deductible amounts: $0, $25,000, $50,000, $100,000, $250,000 and $500,000. Eighteen diagnosis categories, consistent with categories used in the prior large claims study, are analyzed. Data are also presented for claimants with unknown primary diagnosis. Descriptions and ICD-9 ranges for each category are presented in Table III. Separate columns are presented for each study year. The table presents two blocks of data for each deductible amount, summarizing frequency and severity.

The frequency data includes the number of claimants within the primary diagnosis category and having total paid charges exceeding the deductible amount. Also presented for each diagnosis category and year is the percentage of claimants represented, compared to all claimants with total charges exceeding the deductible amount. Each claimant's primary diagnosis category was determined during aggregation of claims. The primary category is the category with the highest subtotal of charges for all claims with an ICD-9 within the category range. A claimant's primary diagnosis category need not correspond to the claimant's primary ICD-9 (having the highest subtotal charges among the ICD-9s represented in that claimants claims). For example, a claimant's highest cost ICD-9 code falls outside the range of the primary diagnosis category if several lower cost ICD-9s fall within the primary category range and total an amount larger than the cost of the primary ICD-9.

The severity data presents paid charges in excess of the deductible amount (i.e., the deductible amount is not included in these subtotals), by diagnosis category and year. Also presented are each category's excess charges as a percentage all claimants' charges which exceed the deductible. Finally, the severity data summarize, by diagnosis category, the average excess charges per claimant having charges in excess of the deductible and having a primary diagnosis falling within the category.

IV. ANALYSIS BY RANGE OF CLAIMANTS' CHARGES AND DEDUCTIBLE: CLAIMANTS AND PAID CHARGES IN RANGE; CLAIMANTS, EXCESS CHARGES AND AVERAGE EXCESS CHARGES ABOVE RANGE MINIMUM; ANNUALIZED SEVERITY TRENDS; NORMALIZED DATA

Table IV presents analysis of claims data by fifty-two ranges of a claimant's total paid charges, for each study year. Presented as separate worksheets are Table IV-A for all claimants, regardless of plan type, and Table IV-B for claimants with a PPO plan type. Separate blocks of information are presented for raw data and normalized data. The current design prints each worksheet as six pages, with each year's raw data page followed by its normalized data page. The user can easily print all raw data before all normalized data by changing the "page order" in "page setup" to "down, then over".

Presented for each year and charge range are the number of claimants with paid charges within the charge range and the subtotal of their paid charges. Computed from these data are columns for the number of claimants with paid charges exceeding the minimum amount of the range and the subtotal paid charges for these claimants (these subtotal, still include amounts below the range minimum). The next column views the charge range minimum as a deductible amount and subtracts it, resulting in an expression of excess charges above the deductible. Average excess charge per claimant having charges exceeding the range minimum is then presented. One-year severity trends are presented for claim years 1998 and 1999 (comparing the prior year). A two year annualized severity trend is presented for claim year 1999 (comparing 1997). The severity trend expresses the percentage change in average excess charge per claimant from the base year to the current year. The annualized two-year severity trend is the annual percentage rate that would have to be compounded for two years to achieve the observed two-year change.

For columns other than average excess charge per claimant and the severity trends, normalized data are presented in a block to the right of the raw data. Claimant count data are presented as a percentage of all claimants, and charge data are expressed as a percentage of all charges.

V. ANALYSIS OF HOSPITAL CHARGES BY RANGE OF CLAIMANTS' TOTAL PAID CHARGES, FOR SUBSET OF INSURERS PROVIDING HOSPITAL CHARGES: CLAIMANTS, TOTAL PAID CHARGES, HOSPITAL PAID CHARGES

Data for six of the seven insurers remaining in the database include hospital paid charges, some analysis of which is summarized in Table V. Hospital charges include both in-patient and out-patient charges. This table analyzes the data by range of claimants' total paid charges. Table V-A and Table V-B, prepared as separate worksheets, present results for all plan types and for PPO plan types, respectively. Each Table is designed to print data for each study year on a separate page.

This analysis was performed on a version of the database, which included insurer identity. To protect insurer identity, claimants for the one insurer not providing hospital paid charges are not flagged in the public release database. Consequently, a user cannot replicate data for "claimants in range" and "total paid charges in range" for the subset of six insurers providing hospital data. Subtotaling charges and claimant counts by range of total paid charges, across all insurers, would include charges and counts for the insurer not providing hospital charges. That insurer's hospital charge data would not be included in hospital charge subtotals because its hospital charges were not provided. Therefore, such analysis would skew the results, underestimating hospital charges as a portion of total charges. For the six insurers providing hospital charges, claimants without hospital charges have blank values (93.4% of such claimants) or values of $0 (6.6% of such claimants) in the hospital charge field. For the three study years combined, 39.4% of the 3,070,190 claimants covered by these six insurers incurred hospital charges.

Analysis for Table V assigns each claimant to a charge range according to the total paid charges (hospital and non-hospital combined) for that claimant. For each total paid charge range and year, Table V presents claimant counts, total charges and hospital charges (in-patient and out-patient hospital charges, combined). Hospital charges as a percentage of total charges are then computed for each total paid charge range. Total charges and hospital charges are computed for all claimants having total paid charges above the minimum of each range (no subtractions are made from these subtotals, these are not excess charges above a deductible amount). Finally, the ratio of hospital charges to total charges is expressed as a percentage, for claimants with total paid charges above the range minimum.

The reader may note a few entries on Table V – particularly for Total Paid Charges between $375,000 and $425,000 – for which the Hospital Charges as a Percent of Total Charges appears unusually low. These records have been examined carefully, and there is no clear evidence that these unusual observations are caused by anything other than random fluctuation.

VI. ANALYSIS OF HOSPITAL CHARGES BY RANGE OF CLAIMANTS' HOSPITAL PAID CHARGES, FOR SUBSET OF INSURERS PROVIDING HOSPITAL CHARGES: CLAIMANTS AND HOSPITAL PAID CHARGES IN RANGE; CLAIMANTS, HOSPITAL CHARGES, EXCESS HOSPITAL CHARGES AND AVERAGE EXCESS HOSPITAL CHARGES ABOVE RANGE MINIMUM

For the subset of six insurers providing hospital paid charges, analysis by range of hospital paid charges is presented in Table VI. Table VI-A and Table VI-B, prepared as separate worksheets, present results for all plan types and for PPO plan types, respectively. Each Table is designed to print data for each study year on a separate page. Because this analysis includes only claimants with positive hospital paid charges, the user can replicate this analysis from the public use database.

Analysis for Table VI assigns each claimant to a charge range according to the hospital paid charges for that claimant. For each hospital paid charge range and year, Table VI presents claimant counts and subtotal hospital paid charges. Claimant counts and hospital paid charges are computed for all claimants having hospital paid charges above the minimum of each range (no subtractions are made from these subtotals). Next, the minimum value of each hospital paid charge range is treated as a deductible amount and is subtracted from hospital paid charges to express excess hospital charges above the range minimum. Finally, average excess hospital charges per claimant having hospital charges above the range minimum is computed.

VII. ANALYSIS OF THIRTY HIGHEST COST DIAGNOSIS CODES (ICD-9)

Table Series VII consists of four tables, each designed as a separate worksheet. Tables VII-A, VII-B and VII-C present rankings of the thirty highest cost diagnosis codes (ICD-9) for all claimants, for claimants with paid charges exceeding $25,000, and for claimants with paid charges exceeding $100,000, respectively. The claims data from all three study years were combined for the rankings. The rankings were obtained by subtotaling claimants' total paid charges for each primary diagnosis code appearing in the database. The thirty highest cost diagnosis codes for the claimant group under consideration were retained for each ranking.

A claimant's primary diagnosis is the ICD-9 code associated with the highest subtotal of paid charges among that claimant's claims. The primary diagnosis codes and claimants' subtotal paid charges for claims coded with the primary diagnosis were determined during aggregation of claims. (In fact, each claimant's secondary and tertiary diagnosis codes and associated charges were also determined during aggregation and are included in the public use database.) Omitted from the rankings are primary diagnosis codes that are blank or have value "000". Each ranking is sorted by ICD-9 code, and descriptive diagnoses are provided. In addition to rank, each ranking presents total paid charges for claimants having the primary diagnosis code, subtotal paid charges for those claimants' claims that are coded with the primary diagnosis, and the number of claimants with the primary diagnosis. Average total paid charges per claimant with the primary diagnosis are also presented for each code.

Table VII-D provides an analysis of high cost diagnosis codes by range of total paid charge. The ranking of Table VII-B, for claimants with paid charges exceeding $25,000, is used for this

analysis. Computations for the analysis include all claimants, regardless of the amount of their paid charges. That is, even though the diagnosis ranking utilized is that for claimants with paid charges greater than $25,000, claimants with paid charges $25,000 or less are included in the subtotals of Table VII-D.

Table VII-D is currently designed to print as thirty pages, one for each high cost ICD-9, sorted by cost rank. Separate data columns are provided for each study year. For each high cost ICD-9 and for nineteen charge ranges, Table VII-D presents the number of claimants having the primary diagnosis and the sum of their total paid charges. Totals and averages across all charge ranges are provided at the bottom of each page. In addition, for each charge range, Table VII-D computes the annualized severity trend in excess charges per claimant above the range minimum, treated as a deductible amount. In arriving at these severity trends, the deductibles are subtracted from the sum of total charges for claimants with paid charges exceeding the range minimum. The severity trends express, as a percentage, the change in average excess charges per claimant above the range minimum as a deductible, compared to the base year. The comparison of 1999 with 1997 presents an "annualized" severity trend, which computes the annual percentage that would have to be compounded to achieve the observed two-year change.

VIII. ANALYSIS BY GENDER, AGE RANGE AND DIAGNOSIS CATEGORY: TOTAL PAID CHARGES AND PERCENT OF CHARGES FOR THE DIAGNOSIS CATEGORY

Table Series VIII is designed as sixteen tables, each occupying a separate worksheet. These tables can be consecutively numbered by printing the entire workbook. Each table presents paid charge subtotals and percentages (as separate data blocks) for nineteen diagnosis categories (including unknown diagnosis) and for eight age ranges (including blank or unknown age). Four tables are presented for all study years combined, and four tables are presented for each of the three study years. Each group of four tables includes separate tables for all genders combined, for females, for males and for unknown genders.

For each table (distinguished by time frame and gender(s)) the upper data block presents subtotal paid charges by diagnosis category (separate rows) and age range (separate columns). Subtotals across all diagnoses for each age range are presented in the bottom row of the data block. Subtotals across all age ranges for each diagnosis category are presented in the rightmost column of the data block.

The lower data block for each table parallels the layout of the upper data block. The percentages reported are computed as paid charges for the time frame, diagnosis category, gender and age range (taken from the corresponding cell in the upper data block for the same table) divided by paid charges for the same time frame and diagnosis category, but combining all genders and age ranges (taken from the rightmost column for the diagnosis category in the upper data block for the table summarizing all genders for the same time frame). That is, the percentage reported represents the gender and age range fraction of all paid charges for a given diagnosis category

and time frame. The rightmost column of the lower data block presents a gender's portion of paid charges for a diagnosis category, combining all age ranges.

IX. EXPOSURE ANALYSIS BY GENDER, AGE RANGE AND DEDUCTIBLE AMOUNT, SUBSET OF FOUR INSURERS PROVIDING MEMBER EXPOSURE DATA: NUMBER EXPOSED, CLAIMANTS AND CHARGES EXCEEDING DEDUCTIBLE, PROPORTION EXCEEDING DEDUCTIBLE, CLAIM COST PER EXPOSURE

For a subset of four insurers providing credible member exposure information, Table IX summarizes a claims analysis by gender and by seven age ranges, for eleven deductible amounts. Each page of the table presents results for a different deductible amount. Results for each study year and for all three study years combined are presented as separate data blocks on each page. Columns presenting data for males are placed to the right of those for females, and age ranges are presented by row.

For each deductible, time frame, gender and age range, Table IX presents the number of members exposed, the number of claimants having paid charges exceeding the deductible amount, and the amount by which the sum of their paid charges exceed the deductible. Amounts in the column headed "Charges Exceeding Deductible" have the deductible amount subtracted from the sum of charges for claimants with charges exceeding the deductible. Table IX then computes the fraction of members with claim costs exceeding the deductible, and the average excess cost per exposure. Results across all age ranges are presented in the bottom row of the data block for each time frame.

For zero deductible amount and age range 0 to 1, Table IX indicates a "Frequency Deductible Exceeded" greater than 100 percent (that is, more claimants than members exposed) and high "Claim Cost per Exposure". These observations may result from coverage of newborns for fewer than twelve member-months during the year of birth. The low frequencies and claim costs

for claimants age 65 or older are explained by a mix of primary coverage. Only insured costs are included in the data; Medicare costs are not included.

Presented as a separate worksheet is a note containing data, which reconcile Table IX with Table II-B. Exposure and claims data are omitted from Table IX when member or claimant gender or age are missing from the data. Table II-B was based on all claimants and members, because results were not separately presented by gender and age. The note to Table IX summarizes the data that were omitted from that table.

Exposure data incorporated into Table IX are not included in the claims database. The number exposed by time frame, gender and age range are summarized in Table XI.

X. CORRELATION OF CLAIMANTS YEAR-TO-YEAR, BY CHARGE RANGE: CHARGES AND CLAIMANTS IN RANGE, FOR YEAR(S) AFTER BASE YEAR

Table Series X presents a correlation of claimants from year-to-year, by charge range. Table X-A presents an analysis for base year 1997 and subsequent study years 1998 and 1999. Table X-B is for base year 1998 and subsequent study year 1999. For base year claimants who are identified as having claims in a subsequent study year, Table Series X summarizes claimant counts and charges for the subsequent year. Each table separates base year claimants by four ranges of paid charges: not exceeding $25,000; greater than $25,000 and not exceeding $50,000; greater than $50,000 and not exceeding $100,000; and greater than $100,000. For each base year charge range, each subsequent study year is analyzed by nineteen charge ranges. Each base year charge range is presented on a separate page. The number of claimants and total paid charges within each base year charge range are presented near the top of each page.

The charges and number of claimants within each subsequent year charge range are presented for claimants who are identified as having base year claims within the base year charge range. The charges are then expressed as a percentage of subsequent year charges summed across all charge ranges. Claimant counts are expressed as a percentage of the number of base year claimants within the base year charge range. The number of claimants appearing on the top line of each data block, having no paid charges, is the number of claimants from the base year and base year charge range who have no claims appearing in the subsequent year's claims. This observation does not indicate whether coverage continued with the insurer. Totals across subsequent year charge ranges are presented near the bottom of each page. The average charge per claimant is computed for claimants who did file claims in the subsequent year.

XI. SUMMARY OF MEMBER EXPOSURE BY GENDER AND AGE RANGE: SUBSET OF FOUR INSURERS PROVIDING MEMBER EXPOSURE

Table XI, combined with the exposure flag field in the database (indicating claimants covered by an insurer who provided credible member exposure data), presents member exposure information necessary to replicate Table II-B and Table IX. For the four insurers providing credible member exposure data, Table XI summarizes member-months of exposure, for each study year, by gender and age range.

APPENDIX A: REQUEST FOR INSURER PARTICIPATION

SOCIETY OF ACTUARIES

475 N. MARTINGALE RD., SUITE 800, SCHAUMBURG, IL 60173-2226 847/706-3500
847/706-3599 *FAX*

November 1, 1999

Dear Health Plan Executive:

The Society of Actuaries ("SoA") would like to have your input for a study of medical expense claims. The only medical expense experience study done by the SoA in recent years has been the "Group Medical Insurance Large Claims Database Collection and Analysis," which was published as a monograph in August of 1997. The SoA is looking to repeat and expand upon this Large Claims Study by conducting a comprehensive study of medical expense experience, involving all claims and exposures from a variety of carriers. The SoA expects to obtain data useful to companies and consultants on the major cost variables that affect the pricing of medical care.

Two separate types of analysis will be done with this data. The SoA has charged the Medical Large Claims Experience Committee with the task of analyzing the factors that affect claim incidence rates and the distribution of claim sizes. The Medical Claims Credibility Project Oversight Group will take on the task of analyzing the credibility that can be assigned by carriers to experience when pricing, rerating, or valuing groups or other blocks of individual business.

The Medical Claims Credibility POG will make available the variance/covariance statistics associated with the components that explain variation in medical care insurance costs from carrier to carrier, from group to group and from individual to individual. These statistics are critical to the use of credibility theory to determine the technically correct amount of credibility any single group, or line of business should receive relative to the underlying expectations of the business it represents. A large body of experience including many carriers, groups, and individuals is required to perform such analyses.

The results of this study will be useful to reinsurers and primary carriers of all types. This letter is to solicit your input regarding both the form of the study and the availability of data for the study.

As was done with the prior study, individual contributions are kept strictly confidential. Data will be contributed to the SoA office and will be identified further only on a coded basis. Only aggregate results will be published and only aggregate results will be seen by the Medical Large Claims Experience Committee and the Medical Claims Credibility POG. The SoA also plans to produce a public use database representing the aggregate data on large claims.

Every carrier's contribution to the study would be greatly appreciated, regardless of the size of the block of business. Attached are two documents: Claims Study Specification Issues and a Claims Study Questionnaire. Please review these documents and respond to the questionnaire, with respect to the form of the study and the availability of data for the study, by December 15, 1999. Please contact either Tom Edwalds, SoA Senior Research Actuary at 847-706-3578 or Jack Luff, SoA Experience Studies Actuary at 847-706-3571 if you have any questions. Thank you.

Sincerely,

William R. Lane, Chairperson
Committee on Health Benefit Systems Research
(402-778-0297)

Anthony J. Houghton, Chairperson
Medical Large Claims Experience Committee
(561-845-0022)

CLAIMS STUDY
SPECIFICATION ISSUES

This document outlines the issues relating to a follow-up to the Large Claims study. These were identified by the Medical Large Claims Experience Committee and the Medical Claims Credibility Project Oversight Group.

Form of Contribution

We are anxious for organizations to contribute and will work with contributors to accept data in any reasonable form consistent with the final specifications. One possibility would be to supply a computer file(s) covering the claims during the time period specified and another file(s) of exposure data. Contributions that have been processed fully or partially into the form of the final specifications would also be accepted.

Period of Contribution

The committee is requesting data for the three (3) calendar years 1996, 1997 and 1998. If it is more convenient to contribute on an underwriting year basis, we would like the three (3) underwriting years beginning in 1996, 1997 and 1998.

Scope of Contribution

In the earlier Large Claims Study, we profiled large claims and large claim rates. Only a few contributors were able to provide the exposure data needed to develop rates. In this study, we intend to profile all claims, both large and small, and claim rates (and need both claims and exposures to do so).

Type of Managed Care

This study is intended to encompass all types of Managed Care plans. However, different types of data may be appropriate for different types of Managed Care. Please let us know in the comment section if there are special data needs for your type of plan.

Charge Data

We recognize that all three (3) of total charges allowed charges and paid charges may not be generally available, and some contributors may be unwilling to provide all this data even if it is available. We do want to proceed using the best data available.

Diagnosis & Procedure Codes

The committee is interested in whether the study should be done on the basis of a single primary diagnosis or whether multiple diagnosis codes should be used. If only one, how should the primary value be determined? If multiple, how many should be provided for and how should the analysis be done? Similarly, should procedure code or codes be captured, and if so, on what basis?

Reinsurance & Reinsurance Organizations

The committee wants to proceed with this study in a manner that gives proper recognition to reinsurance with respect to health claims. The committee also wants to collect data that would produce results that would be useful to organizations doing reinsurance with respect to health claims.

Contribution Media

Although this will be affected by the amount of data contributed, the committee is interested in any issues relating to the physical media used to submit contributions.

SOCIETY OF ACTUARIES
CLAIMS STUDY
QUESTIONNAIRE

(Please respond by December 15, 1999)

Company Name:_____

1. Is your company willing to provide data to a claims study?

___ Yes
___ No

If no, why not?_____

_____
(skip to Question 12)

2. Can your company provide a three-year block of data (nominally 1996, 1997 & 1998)?

___ Yes
___ No

If no, identify the years for which data could be provided_____

3. Would you prefer to contribute on a calendar or contract year basis?

___ Calendar
___ Contract (please describe)_____

_____

4. On what basis can your company provide data?

___ All claims plus exposure data
___ Large claims plus exposure data
___ All claims only
___ Large claims only

5. Please indicate which of the following your company could contribute to our study: (check all that apply)

___ Total charges ___ Allowed charges ___ Paid charges

6. Would your company prefer to contribute data on the basis of a single primary diagnosis or should multiple diagnosis codes be used?

___ Single ___ Multiple How many? ____

What basis should be used?_____

_____

7. Does your company feel that procedure code(s) should be captured and, if so, on what basis?

___ Yes (single) ___ Yes (multiple) How many? ____
___ No

If yes, what basis should be used?_____

_____

8. With respect to media form (magnetic tape, tape cartridge, diskettes, etc.), what would be the most convenient way for your company to contribute data?

Claims_____

Exposures_____

9. Is date of birth data available for all claimants?

___ Yes
___ No, only for covered employees
___ No, only age is available

10. Is the zip code of the claimant available?

___ Yes
___ No, only the zip code of the covered employee

11. How should the categorization of charges into Hospital, Physician and Other be done? Should it be by procedure code? Should the final specifications include guidance in this area?

12. Other comments?

_____

_____

_____

_____

_____

_____

Completed by: _____

Title: _____

Phone: _____

E-Mail: _____

FAX: _____

Contact Person (if not person completing form):

Name: _____

Title: _____

Phone: _____

E-Mail: _____

FAX: _____

Please return the completed form by December 15, 1999 to:

Jack Luff
Experience Studies Actuary
Society of Actuaries
475 N. Martingale Road, Suite 800
Schaumburg, IL 60173-2226

Phone: 847-706-3571
Fax: 847-706-3599
E-Mail: jluff@soa.org

Thank you for your support of Society of Actuaries Studies.

APPENDIX B: DATA SPECIFICATION

SOCIETY OF ACTUARIES
475 N. MARTINGALE RD., SUITE 800, SCHAUMBURG, IL 60173-2226 847/706-3500

847/706-3599 *FAX*

Claims Study
Data Specifications

The data request is organized along a "block – exposure – claim" framework.

We define a "block" of business to be a group of policies with similar marketing, pricing, and underwriting characteristics. Many contributors may have only one block. Other contributors may have separate lines of major medical business (e.g., individual, small group reformed business, fully insured group, stop loss) or may have purchased blocks of business. We ask the contributor to use it's own block identification scheme. Contributors are welcome to add codes to identify blocks that do not fit the scheme we have outlined.

We are asking for individual exposure records for each primary insured person, and if coverage changes for any such person, separate records for each change. A "primary insured person" is the certificate-holder in a group plan, or the policyholder in an individual plan. Contributors may provide us with a plan-code type field in place of the series of benefit description fields (items 8 through 21) in the exposure record. In that case, the contributor should provide us with a plan code table, showing us how items 8 through 21 should be completed for each plan code.

For claims information, we are asking for individual records for each claim encounter whose service date falls in the three-year study period, with each claim record tying back to an exposure record.

<u>BLOCK DEFINITION</u>

1. *Block name*. This will be referred to in the exposure records.
2. *Business type:* (A) individual, (B) association of individuals, (C) association of groups, (D) small group, (E) large group, (F) Taft-Hartley
3. *Nature of block:* (A) open block, new business rates in line with renewal rates, (B) open block, new business rates substantially lower than renewal rates, (C) open block, full community rating, (D) closed block, potentially in assessment spiral, (E) closed block, full community rating, (F) other closed block. [NB: Some of these categories are more appropriate for individual business than for group business.]
4. *Underwriting:* (A) full medical underwriting, (B) limited medical underwriting, (C) no medical underwriting except possibly for late enrollees
5. *Handling of pre-existing conditions:* (A) both riders and rating, (B) riders only, (C) rating only, (D) none

EXPOSURE RECORD

1. *ID# of primary unit* . This would be the group certificate number for group coverage, or the policy number for individual coverage.
2. *ID# of primary insured* . The primary insured is the employee for group coverage, or the policyholder for individual coverage.
3. *Block name.* This will tie the exposure record to one of the blocks defined above.
4. *Original issue date of group certificate / individual policy.*
5. *Coverage effective date.* [NB: For groups/individuals insured through the entire period of the study, the coverage effective & termination dates (also see item 6) would be the endpoints of the study period; otherwise, they would reflect the subinterval of the study period over which coverage applied. If elements of the insured's plan design were to change while the insured remained covered, e.g. at policy renewal, then a new exposure record should be generated.]
6. *Coverage termination date.*
7. *SIC Code*.[NB: This applies only for group coverage.]
8. *Type of Managed Care:* (A) indemnity without UR, (B) indemnity with UR, (C) PPO panel rented from an HMO, (D) other PPO, (E) POS, (F) IPA HMO, (G) staff model HMO, (H) other HMO
9. *In-network deductible.*
10. *Out-of-network deductible.*
11. *Separate deductible for prescription drugs:* no, or specify
12. *In-network coinsurance band.*
13. *Out-of-network coinsurance band.*
14. *In-network out-of-pocket limit.*
15. *Out-of-network out-of-pocket limit.*
16. *Prescription drug card:* (A) drug card, claims cannot be tied back to claimant, (B) drug card, claims can be tied back to claimant, (C) no drug card
17. *Physician encounter fee / copay:* no, or specify
18. *Prescription drug copay:* no, or specify
19. *Limit on Maternity Benefits*: not covered, or SAAO, or specify limit
20. *Limit on Mental/Nervous Benefits:* SAAO, or specify limit
21. *Limit on Drug/Alcohol Benefits:* SAAO, or specify limit
22. *ZIP Code.*
23. *Type of coverage:* (A) single only, (B) single + spouse, (C) single + child(ren), (D) family, (E) child(ren) only
24. *Continuation/conversion:* (A) COBRA-type continuation, (B) conversion from a group policy, (C) neither, (D) cannot identify
25. *Birthdate of primary insured.*
26. *Sex of primary insured.*
27. *Smoking status of primary insured.*
28. repeat trailers 25-27 for each dependent, adding a field to explain relationship of dependent to primary insured (self/spouse/child)

CLAIM RECORD

1. *ID# of primary unit.*
2. *ID# of primary insured.* These two fields tie the claim record back to the exposure record.
3. *Birthdate of claimant.*
4. *Sex of claimant.*
5. *Relationship of claimant to insured*: (A) self, (B) spouse, (C) child
6. *Date of claim service encounter.*
7. *Covered charges for hospital services.*
8. *Allowed charges for hospital services.*
9. *Paid charges for hospital services.*
10. Repeat trailers 7-9 for physician services.
11. Repeat trailers 7-9 for other services.
12. *Primary diagnosis or diagnoses*
13. Primary procedure code.

APPENDIX C: PUBLIC USE DATABASE

The public use claims database is provided as three ASCII, delimited files, one for each study year. Fields are separated by commas; text fields are delimited by double quotes; numeric fields are not delimited. The first line of each file contains field names. The database size in its ASCII format, for the three files combined, is approximately 690 megabytes.

The following table lists each field of the database, in order, including the field names assigned on the first line of each file, a field description, and a recommended data type.

Record Structure

No.	Name	Description	Type
1	CLAIMYR	Year claims were paid	A4
2	CLAIMANT	Unique identifier for each claimant	N
3	RELATION	Relationship to Subscriber (E=Subscriber, S=Spouse, D=Dependent)	A1
4	PATSEX	Gender of Claimant (M or F)	A1
5	PATBRTYR	Birth Year of Claimant	A4
6	HOSCVCHG	Covered Hospital Charges	N
7	HOSLWCHG	Allowed Hospital Charges	N
8	HOSPDCHG	Paid Hospital Charges	N
9	PHYCVCHG	Covered Physician Charges	N
10	PHYLWCHG	Allowed Physician Charges	N
11	PHYPDCHG	Paid Physician Charges	N
12	OTHCVCHG	Covered Other Charges	N
13	OTHLWCHG	Allowed Other Charges	N
14	OTHPDCHG	Paid Other Charges	N
15	TOTCVCHG	Covered Total Charges	N
16	TOTLWCHG	Allowed Total Charges	N
17	TOTPDCHG	Paid Total Charges	N
18	DIAG1	Diagnosis with Highest Subtotal of Paid Charges	A3
19	DIAG1CHG	Subtotal of Paid Charges for Highest Cost Diagnosis	N
20	DIAG2	Diagnosis with Second Highest Subtotal of Paid Charges	A3
21	DIAG2CHG	Subtotal of Paid Charges for Second Highest Cost Diagnosis	N
22	DIAG3	Diagnosis with Third Highest Subtotal of Paid Charges	A3
23	DIAG3CHG	Subtotal of Paid Charges for Third Highest Cost Diagnosis	N
24	DGCAT	Diagnosis Category with Highest Subtotal of Paid Charges	A33
25	DGCATCHG	Subtotal of Paid Charges for Highest Cost Diagnosis Category	N
26	EXPOSMEM	Flag field: “Y” if included in member exposure, “N” otherwise	A1
27	PPO	Flag field: “Y” if covered by PPO, “N” otherwise	A1

Type “An” indicates an alphanumeric (i.e., text or character) field, the values of which have a maximum length or number of characters, “n”. Type “N” indicates a numeric field.

Each record or row of the database presents a summary of claims data for an individual claimant for one year. This summary was derived from the claims level data (i.e., a separate record for each claim) contained in the much larger source data sets provided by each insurer. The annual claims of each claimant were aggregated to one record, subtotaling charges and identifying primary diagnosis codes and categories.

For each claimant, paid charges were subtotaled by diagnosis code. These subtotals were then ranked by amount. The three highest subtotal charges and their codes were included in the database. In the event of equal subtotal charge amounts, the codes were selected by their alphanumeric sort order.

In addition, based on diagnosis code, the diagnosis category was appended to each record at the claim level, and charges were subtotaled by the resulting diagnosis category. Each claimant's highest subtotal charge and its associated diagnosis category were included in the database. ICD-9 ranges used to determine the diagnosis category, consistent with the prior large medical claims study, are presented in the following table:

Diagnosis Category Definitions

ICD9MIN	ICD9MAX	DIAGNOSIS CATEGORY
001	139	Infectious & Parasitic Disease
140	239	Malignant Neoplasms
240	279	Endocrine & Metabolic Disorders
280	289	Blood Related Disorders
290	319	Mental Disorders, Drug, Alcohol
320	359	Nervous System
360	389	Sense Organs
390	459	Circulatory System
460	519	Respiratory System
520	579	Digestive System
580	629	Genitourinary System
630	679	Pregnancy & Childbirth
680	709	Skin Disorders
710	739	Skeleton & Muscle System
740	779	Congenital & Perinatal
780	799	Symptoms & Ill-Defined Conditions
800	999	Injury & Poisoning
V00	V84	Health Status or Service

The number of claimants and total paid charges in the database, by year, are summarized in the following table:

Year	Number of Claimants	Total of Paid Charges
1997	1,241,438	2,003,162,217.76
1998	1,460,854	2,466,093,740.87
1999	1,591,738	2,599,356,657.86
All Years	4,294,030	7,068,612,616.49