Paper 253

Using SAS® Software to Build a Data Warehouse of Standardized Medical Costs and Patient Utilization for Health Services Research

Priscilla Van Grevenhof, Mayo Clinic, Rochester, MN
Judith L. Wagner, Congressional Budget Office, Washington, DC
Eva R. Helgeson, Mayo Clinic, Rochester, MN
James M. Naessens, Mayo Clinic, Rochester, MN
Jill M. Killian, Mayo Clinic, Rochester, MN
and Roger W. Evans, Mayo Clinic, Rochester, MN

INTRODUCTION

A collaboration of researchers at the Mayo Medical Center and Olmsted Medical Center created a SAS® data warehouse of detailed billing data. This unique warehouse contains billing data for most of Olmsted County from 1987 to present. The challenge for the developers was to create a warehouse of all historical data stored on different platforms and in a variety of different forms from five different institutions. The goal is to give researchers easily accessible, consistent data for analysis. SAS tools were used in the development of the warehouse and are used for accessing, manipulating and analyzing data.

OBJECTIVES

As an adjunct to the Rochester Epidemiology Project, researchers from the Mayo Clinic and the Olmsted Medical Center have collaborated to create a SAS data warehouse for population-based studies of health care utilization and costs. Detailed billing data, representing the vast majority of medical care services for residents of Olmsted County, Minnesota, are available since 1987. Because most research questions have relevance beyond the boundaries of Olmsted County, the costing methodology, when applied to the detailed utilization, provides a standard unit cost measure that is adjusted for both inflation to a base year and geographic variation. The methodology used protects proprietary cost data from disclosure and will provide consistent unit cost estimates for ongoing research studies of health care costs.

PROJECT LIFECYCLE

A multidisciplinary group of researchers, epidemiologists, statisticians, economists, and Information Services staff was formed to gather information about needs and expectations of the Cost/Utilization Data Warehouse. This group decided the data warehouse development should be in SAS for the following reasons: (1) in-house programmer expertise, (2) statistical analysis expertise, (3) knowledge of SAS tools for interfacing and accessing source systems and (4) SAS availability on four platforms at Mayo Clinic. The Cost/Utilization Data Warehouse project group formed a working group of five people from the above disciplines. The process of development consisted of the following six distinct tasks.

DEFINE CORE DATA ELEMENTS

The working group's first issue was to determine the feasibility of obtaining the selected set of data elements to accomplish the objectives of the project. To gather this information, the work group had to do the following:

- Review systems in the marketplace attempting to accomplish some of the same goals;
- Determine the data sources available to draw together the information;
- Determine what was not available at all, what was available only during specific time frames, and what futures were planned in the source systems;
- Determine how to map all data sources to a common set of data elements with consistent definitions, codes and formats across all years; and
- Document the core data elements and the limitations found during the information gathering.

INSURE DATA QUALITY

The issue of data quality was confounded by the number of systems used to build the warehouse. The same data could be found in several systems. We wanted the best sources for the data and had to determine system by system, which source to use. It was necessary to develop processes for data scrubbing before the standards were applied. A process of editing, cleansing, verifying, removing duplicates, and determining gaps in the data was developed for each system. If gaps were found in the data, additional source systems were identified to aid in a complete data capture.

STANDARDIZE DATA

With the common set of data elements chosen, reference tables containing codes, associated coding system descriptions and coding translation tables were developed. Tables were also used to drive the costing methodology such as Medicare fee schedules and hospital cost reports. Over 40 tables were created to assist in the effort of standardization of data over time, systems and institutions.

APPLY COSTING METHODS

Applying a standardized, inflation-adjusted estimate of the cost of each service and procedure, reflecting the national average of providing the service in constant 1995 dollars, using Medicare hospital cost reports and fee schedules was the basis for the costing methodology. The approach was tested on random samples of each year and preformed on each source of data. Our basic cost algorithm was also tested on a group of high cost items. This process identified some additional work that needed to be done in mapping CPT4 Codes and UB92 codes. Once the fundamental methodology was verified, it was applied to the entire data warehouse.

VERIFY PROCESS AND DATA

Data verification was accomplished in several ways. A review of aggregate values grouped by year and site, served to identify patterns within the data that would not otherwise be evident. Summary statistics sometimes forced us to re-examine the data in which critical errors were found. It was an unexpected pattern in one site's costs that led us to discover several miscoded codes in the source data, significantly affecting our estimated cost calculations in 1987 and 1988. A randomly selected group of line items from each site, year and for each type of cost algorithm were reviewed. Each step of our process of applying costs, standardization, and data scrubbing was manually verified in those samples.

Finally, abstracters completed chart reviews on all line items for a sample of patients in three disease groups. They compared medical records to line items associated with an individual in the data warehouse and found that an overwhelming majority of information was correctly captured.

PUBLISH METADATA

Metadata turns raw data into usable knowledge and is the documentation of the warehouse. The intent of our Intranet-based metadata is to present three types of information.

- Technical encompasses sources of data, methods and rules of transformation, and data flow for each source in the warehouse.
- Business includes data definitions, summary tables, how to access, how to use the warehouse and the 40 descriptive and conversion reference tables.
- Analytical describes interpolation methods, cost algorithms, aggregate measures by time period and site of service and the 15 tables used to determine costs.

PROBLEMS

Mapping different systems into a common format is a laborious task. Experts in each of the systems are used to assist in re-mapping. This problem will continue as we move to add each new year into the warehouse. When re-mapping of procedural codes is not possible, an interpolation method is used to cost out services. Less than 3% of the services required the interpolation method.

Documentation is essential to the success of this system. We were not prepared for the amount of time this effort required and continues to require. Using the Intranet permitted us to eliminate manuals and the problems of updating them.

CONCLUSION

The initial projects using the SAS data warehouse examine patterns of care for disease-specific populations. The warehouse concept offers researchers a robust resource delivering the data in an easy-to-use, consistent, and accessible format. Any provider/payer with an adequate historical billing archive and processing system could potentially replicate our process. Our community-wide population-based medical care costing resource is a valuable tool for investigation of many issues relevant to health services research.

Measuring differences in health care costs attributable to patient characteristics, interventions, or temporal trends has become an integral part of health care outcomes research. Studies of relative costs or cost-effectiveness are increasingly relevant to clinical decision making. In all such studies, the objective is either to compare the value of health care resources used or estimate the effect of patient or intervention characteristics on the cost of health care. With the standard costing methodology used by this system researchers now have the ability to make these comparisons.

TRADEMARKS

SAS is a registered trademark or trademark of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

REFERENCES

Anderson, Scott and Morton, Steve. *Data Modeling for SAS Data Warehouse*. **A SAS Institute White Paper.**

Emmerich, Thomas. *The Rapid Warehousing Methodology*. **A SAS Institute White Paper.**

Inmon, Bill, Imhoff, Claudia, and Gosselin, Larry. *Managing The Data Warehousing Effort: Advanced Topics*. **Barnett Data Systems Course Notes.**

SAS Institute Inc. (1996), *The SAS Data Warehouse.* **A SAS Institute White Paper.**

Tidename, Sue and Chu, Robert. *Building Efficient Data Warehouses: Understanding the Issues of Data Summarization and Partitioning*. **Proceedings of the 21st Annual SAS Users Group International Conference**.

CONTACT INFORMATION

Contact the author at:

Priscilla Van Grevenhof
Mayo Clinic
200 First Street, SW
Rochester, MN 55905
Phone: (507)284-5585
Fax: (507)284-1731
e-mail: vangreve@mayo.edu