Imputation of financial fields in Integrated Anonymous Patient-Level Databases

EVGUENIA JILINSKAIA
PharMetrics, Inc., 311 Arsenal street, Watertown, Massachusetts, 02472
ev@pharmetrics.com

TERI CONDON GOODWIN
PharMetrics, Inc., 311 Arsenal street, Watertown, Massachusetts, 02472
tcondon@pharmetrics.com

CATHY JOHNSON
PharMetrics, Inc., 311 Arsenal street, Watertown, Massachusetts, 02472
cjohnson@pharmetrics.com

STANLEY NORTON
PharMetrics, Inc., 311 Arsenal street, Watertown, Massachusetts, 02472
snorton@pharmetrics.com

Abstract

Rate based analyses such as per member per month (PMPM) and units of measure per 1,000 are important for understanding cost and utilization of a given population. While retrospective claims data is rich with information regarding utilizers of health care services (patients), it does not provide the necessary information for population based analyses. For population based analyses one must have enrollment data as well, which often is not as readily available. This paper describes the process by which PharMetrics, Inc. determines a total population in the absence of detailed, high quality, enrollment data.

Keywords: imputation, exponential predictive model, enrollment data, rate based analyses

1. Introduction

The PharMetrics Integrated Outcomes Database (POID) constitutes an integrated set of fully adjudicated medical and pharmaceutical claims for all covered services. It includes both inpatient and outpatient diagnoses and procedures, and both standard and mail order prescription records. It contains claims records from over 24 million lives belonging to 36 national and regional managed care organizations. The records are representative of the national managed care population on a variety of demographic measures including: geography, age, gender and product type. The data are also longitudinal, with an average member enrollment time of two years. The breadth and depth of the POID allows for comparisons across a range of patient demographics, including age, gender, treatment patterns and co-morbidities.

Currently, of the 36 health plans that contribute claims data to the POID, 60% also contribute enrollment data. All rate-based calculations, such as costs per member per month (pmpm), rates per thousand and prevalence, depend upon quality member level enrollment history data. Consequently, it was necessary for PharMetrics to develop a method for determining population whether or not enrollment data was contributed. Additionally, a method for determining the quality of the submitted enrollment data was necessary.

To evaluate both the quality issues and overcome the absence of enrollment information, the following process of enrollment imputation was developed:

1. The quality of enrollment and claims data is determined based on exponential predictive model for percent of enrollees without claims, as a function of total number of enrollment months within a given year. The quality of enrollment within new incoming data is determined based on the value of the R-square.[2] Regression equations with R-square > 0.8 reflects the decision to include new files for these years into a valid ``learning sample" of enrollment files of sufficient quality.

2. Based on the ``learning sample", values of enrollment weights for re-calculating N of enrollees from number of claimants are estimated separately for each of 20 defined age-gender demographic groups. The end result is a reference table of enrollment weights that can be applied against claims data.

3. Based on the ``learning sample", values of the average number of member enrollment months per year are estimated for each of 20 age-gender demographic groups. The total number of member months for health plans with incomplete or non-existent enrollment data is estimated, summarizing counts of patients in each of the demographic categories using the reference table of enrollment weights and enrollment duration for each health plan, by year. The proposed method of enrollment imputation was validated against the POID for health plans with complete, high-quality enrollment data. It showed stable results and was implemented for PMPM calculations in a Rhinitus study. It will also be used for rate-based analyses in the future.

Definitions for terms used throughout this paper are:

- **Enrollee/Member** – Individual enrolled and eligible for coverage in a health plan
- **Claimant/Patient** – Enrollee who has received health care services for which the provider has been reimbursed.
- **Enrolled Months/Duration** – Number of months each member is enrolled.
- **Member Months** – Sum of the number of months each member is enrolled. It is the denominator for rate based analysis.

2. Exponential Model

Quality of enrollment data is determined based on an exponential predictive model for the percent of enrollees without claims as a function of the total number of enrollment months within a give year. Normally, analysis focuses on the 'utilizers' of health care services. In order to determine the entire population, it's the inverse, in that one must determine the number of 'non-utilizers.'

The process for checking quality of enrollment files is based on a very simple idea:

The probability, that the patient has at least one claim P1, increases with a longer enrollment duration (consequently the probability, that the patient has no claims P0=1-P1 decreases).

Let us denote y the ratio of enrollees without claims to the total number of enrolled people, with x the number of months of enrolment during given period of time. Assign $0 < i <= 12$ (twelve consecutive months are not necessarily in the same calendar year). As it has been found in the study, generally the good correspondence of enrollment and claims data result in a good fit to the following exponential regression equation:

$$y = \theta_1 \exp(-\theta_2 x)$$

For each health plan that submitted enrollment data, a regression equation was determined. Each plan with an R-square greater than .8 was included as part of the learning sample.

Figure 1 shows an example of such a fit for six different health plans with R-squared in the range of 0.88 - 0.983

Figure 1

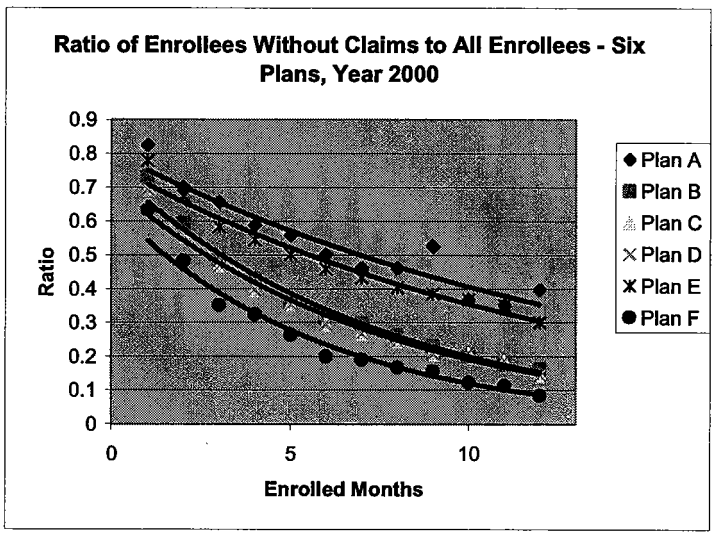
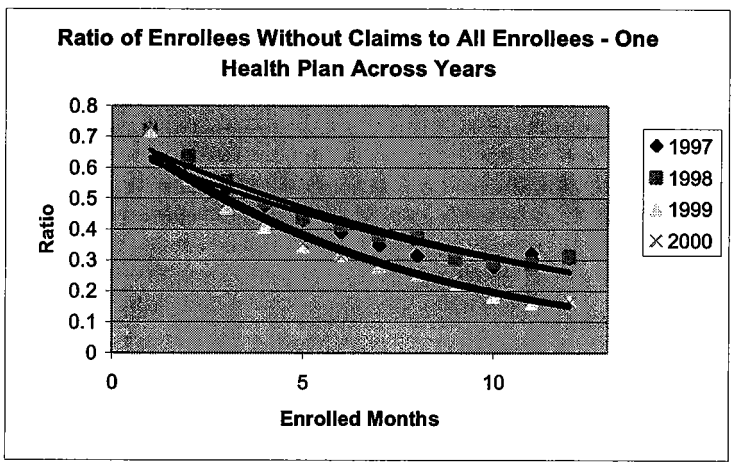


Figure 2 shows an example of such a fit for one health plans across four years with R-squared in the range of 0.93 - 0.95

Figure 2



3. Demographics of Enrolled Population

Data was divided into 20 demographical clusters for each year of data, each mutually exclusive, and in sum exhaustive of the whole population for each year. PharMetrics defined the following clusters:

Males under 3 years of age	Females under 3 years of age
Males 3 to 5 years of age	Females 3 to 5 years of age
Males 6 to 11 years of age	Females 6 to 11 years of age
Males 12 to 17 years of age	Females 12 to 17 years of age
Males 18 to 24 years of age	Females 18 to 24 years of age
Males 25 to 34 years of age	Females 25 to 34 years of age
Males 35 to 44 years of age	Females 35 to 44 years of age
Males 45 to 54 years of age	Females 45 to 54 years of age
Males 55 to 64 years of age	Females 55 to 64 years of age
Males 65 years of age and over	Females 65 years of age and over

Frequencies of each cluster were compared with numbers for targeted population, taken from the U.S. Census Bureau's web-site and the distributions were found to be comparable. [1]

4. Results

When only claims data are present, there are two pieces of information missing that are necessary to determine the population – the number of all enrollees and the period of enrollment for each enrollee. Therefore, these are the figures that must be imputed. Figures 3 and 4 show the mean months of duration, by demographic clusters for the learning sub-sample.

Figure 3

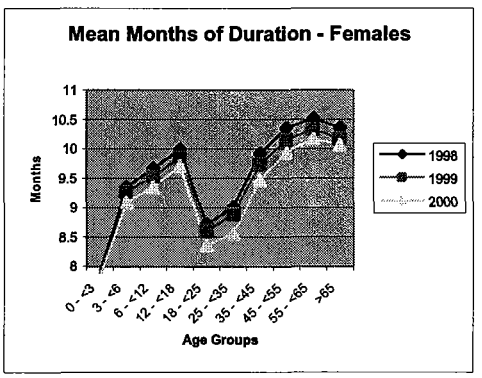
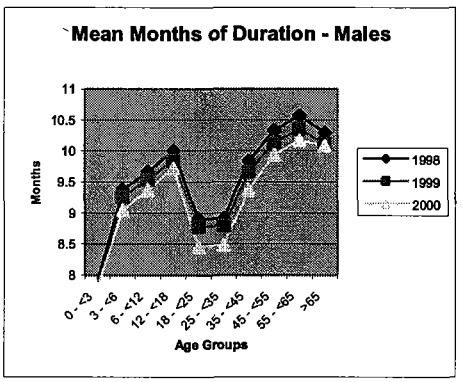


Figure 4



In order to determine the second missing variable, the learning sub-sample was used to define the ratio of data present in claims (the number of patients) to impute the data

that is not present in claims (number of enrollees). The imputed variable is called 'Enrollment Weight'. For each year, 1998, 1999 and 2000, the formula used to determine the enrollment weight (W) was:

$$W_j = M_j / N_j$$

Where $_j$ represents the demographic clusters, M represents the number of enrollees and N represents the number of claimants. Figures 5 and 6 reflect the mean enrollment weights, by year, by demographic cluster of the learning sample.

Figure 5

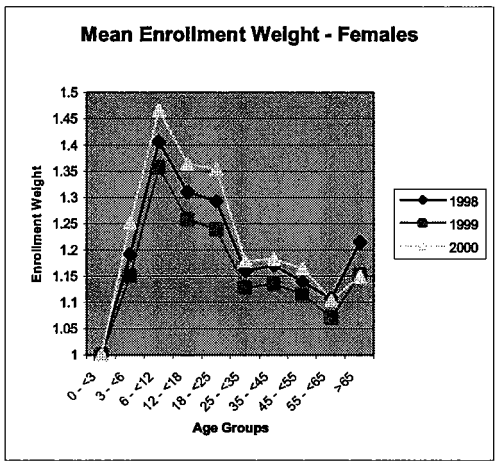
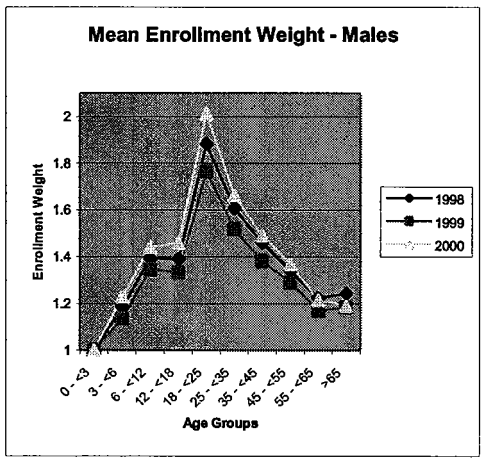


Figure 6



Once both missing variables – duration and enrollment weight – have been imputed, they are applied in the calculation of member months, which provides the population denominator for all rate based analyses with this formula:

$$denom_A = \sum_j n_j * W_j * duration_j$$

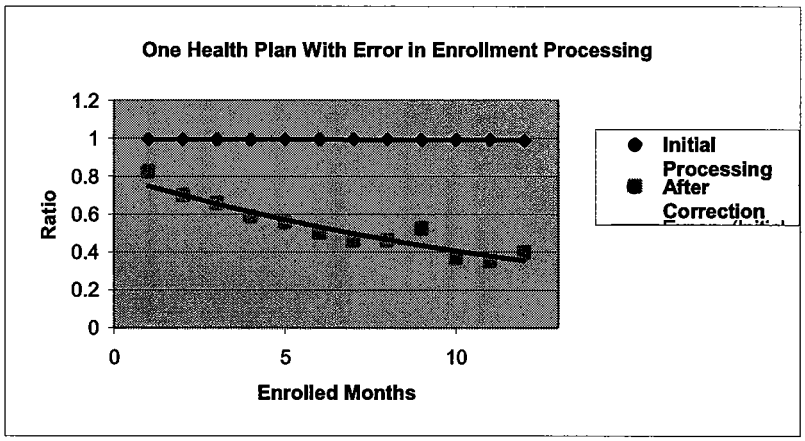
The annualized denominator is the sum of the number of claimants in each group (n) times the enrollment weight (W) for each demographic cluster, times the duration factor for each cluster. In the event the data covers a fraction of the year (e.g., 1st and 2nd Quarter of 2001), the formula can be de-annualized by multiplying the annualized denominator by the fraction (t/12) where t is the number of months of available data.

5. Application in Determination of Quality of Enrollment Data

Data was received from a health plan over two iterations. The first included claims from the time period of January 1999 through December 2000, the second covered the time period of January through June

2001. When data are received this way PharMetrics employs a process of joining the two submissions together. When the data quality method (regression equation) of the enrollment data was performed, it was clear that a problem existed (Figure 7). Further analysis revealed that a processing error had been made and that the joining had not taken place. Therefore, the method was useful in correcting a processing error.

Figure 7



7. Conclusion

The quality and volume of the enrollment and claims data in the PharMetrics Integrated Outcomes Database lends itself to developing a robust and sound method for determining the quality of enrollment data and creating a substantial learning sample. Additionally, rate based analyses can be reliably performed using the PharMetrics method of imputation of membership denominators.

References

[1] http://www.census.gov/hhes/hlthins/historic/.
Table HI-2. Health Insurance Coverage Status and Type of Coverage--All People by Age and Sex: 1987 to 1999
Table HI-3. Health Insurance Coverage Status and Type of Coverage—Children Under 18 by Age: 1987 to 2000

[2] R.J.Larsen, M.L.Marx, An Introduction to Mathematical statistics and its applications, Prentice Hall, 1986

[3] R.J.A. Little, D. B. Rubin, Statistical Analysis With Missing Data, Wiley Series in Probability and Mathematical Statistics, New York,1987.