Sampling a Large Health Care Database, based on Fractal Nature of Claims Data.

Evguenia Jilinskaia, Cathy Johnson, Stanley Norton,
Chris Amendola, Robert St.John, Bo Chong
PharMetrics, Inc. 150 Coolidge Avenue
Watertown, Massachusetts 02472
USA
ev@pharmetrics.com http://www.pharmetrics.com

January 20, 2002

Abstract

High dimensionality and the extremely large volume of data in Pharmetrics Health Care observational database restricts our ability to implement contemporary advanced statistical analysis. The large volume of data leads to the problem of over-powered statistical analysis, where every difference between groups becomes significant due to large sample sizes. The possibility of creating a sample database to build a robust, yet small, replica of the full production database is explored. Based on the assumption of the fractal nature of data, the behavior of different types of systematic samples are compared to stratified random sampling. It was shown that the patient-level α% systematic sample appeared to be self-similar to the original sample, and the use of 10% (or 5%) systematic samples is statistically sound for projecting query results to the whole database and further projection to the population of interest.

Keywords: Systematic sample, random sampling, fractal structure, Large Volume Database , Bootstrap, Projection .

1 Introduction

This paper studies a healthcare database which gives us an example of a non-probability sample of a very large size, which is restricted to an accessible part of the population. This can lead to the unbalanced representation of different strata of the population.

For example, when viewed by payer type (e.g. Medicare Risk, Commercial, Medicaid), some groups of the population may appear to be underrepresented as there is a lower proportion of patients with Medicare Risk in the database than in the general population. One possible approach to restoring the balance and creating a representative balanced sample is to use a variety of statistical re-sampling techniques. High dimensionality and extremely large volume of data restricts our ability to implement contemporary advanced statistical analysis. Moreover, the large volume of data leads to the problem of over-powered statistical analysis, where every difference between groups becomes significant due to large sample sizes.

PharMetrics decided to explore the effectiveness and accuracy of a subset of the full database. The results show that a systematic, patient level sample retained the appearance of the full database. We also determined that a 10% sample is statistically sound for projections and this and smaller size creates the ability to perform more complex statistical methods.

In addition to the operational advantages of a significantly smaller database for purposes such as querying the data, it allows for more complicated statistical methods, such as:

a) establishing an algorithm for assigning values of probabilities for re-sampling and identifying the minimum sample size necessary. This, in turn, allows us to address the issue of unbalanced representation of some groups of the population.

b) investigating the possibility of utilizing data from different sources in a longitudinal study (on a quarterly or semi-annually basis).

c) estimating the stability of the incidence of rare events, an algorithm to force patients with very rare diseases into derived samples.

d) estimating precision of a sampling procedure, which can be judged by examining the frequency distribution of the estimator if the procedure is applied repeatedly to the same population. [8]

2 The Problem

This work employed PharMetric's integrated patient-level database, which contains anonymized, longitudinal, merged medical and pharmaceutical claims data on 25 Million Americans. Delivery of disease-specific products derived from episode-based clinical and financial data includes regional and national normative analyses using episode- and population- based measures, cost and utilization estimates by disease, treatment patterns and effectiveness, pharmacoeconomics and medical outcomes.

The number of query requests has increased greatly over the last year. Additionally, the queries have become far more sophisticated. As a result, many of them result in multiple passes through the data, taking too long and placing heavy demands on computers.

We decided to explore the possibility of creating a sample database specifically for querying. Our goal was to build a robust, yet small, replica of the production data that was statistically sound for projecting query results, i.e. obtaining counts of diagnoses or medical procedures. These techniques are not applicable to the detailed outcomes and pharmacoeconomics analyses for which we employ the whole database.

The purpose of this paper is to describe our findings and to outline a process for creating and putting into operation a sample production database.

3 Methods

Methods of sampling analysis were used. The selection of methods was based on a combination of goals, such as robustness, consistency, computational efficiency and general applicability.

The following steps were undertaken to test the sample:

1. Compare simple statistics of a 5% and 10% sample dataset from health plans of different sizes to the entire datasets for those plans.

2. Compare the results of the production reports created for the 10% sample of the data from all 36 health plans in the PharMetrics Integrated Outcomes Database.

3. Compare the results of a variety of queries against the 10% sample to the same queries against the full database.

4 Results

Every 20th or 10th patient and all of their claims from 4 health plans were included in the sample. Three of the plans are very small and the last very large. We checked the distribution of gender and record type of the samples and full file.

Results of comparing systematic samples

Both the 5% and 10% samples maintained the same distribution for all metrics as the full production dataset. Although a 5% sample is probably sufficient, we decided that a 10% sample would be most desirable. Certain criteria requested in queries could result in relatively small 'hits', such as newer drugs, long continuous enrollment criteria or rare diseases. Our goal is to maximize efficiency, but keep as many patients as possible so that greater than 80% of all queries will result in sufficient findings and it will not be necessary to query the full production database.

Table 1 shows the distribution of two variables, gender and record type, for the full database and both the 5% and 10% samples. The distribution for the full dataset and the 10% sample are very similar.

Since so many queries have continuous enrollment criteria, we also tested the distribution of enrollment using the sampled patients from the claims data. The metric we chose to test is longitudinality. Each of the datasets included in the evaluation has true enrollment. Bear in mind that detailed enrollment files have both claimants and non-claimants. As a result, we expected to retain at least 5% or 10% of the enrollees in the categories of continuous enrollment greater than 2 years, but we expected that we would have less than 5% or 10% of the enrollees in the categories of enrollment less than two years. Obviously, the longer one is enrolled, the more likely one is to be a claimant. As you can see in table 2 below, distribution in the 10% sample is comparable to the full dataset.

Table 1: Percentage of Gender and Record Type for Four Health Plans
Comparison for full, 5% and 10% datasets

Dataset	Metric	Full Dataset	5% Sample Dataset	10% Sample Dataset
A	Gender - Female	61.36	64.30	62.91
A	Gender - Male	38.64	35.70	37.09
A	Record Type - Ancillary	36.59	35.92	36.41
A	Record Type - Facility	0.32	0.31	0.31
A	Record Type - Management	23.54	23.68	23.53
A	Record Type - Pharmacy	37.45	37.96	37.65
A	Record Type - Surgical	2.10	2.13	2.11
B	Gender - Female	69.51	70.48	70.42
B	Gender - Male	30.49	29.52	29.58
B	Record Type - Ancillary	44.35	43.31	43.82
B	Record Type - Facility	0.32	0.35	0.35
B	Record Type - Management	22.38	23.18	22.83
B	Record Type - Pharmacy	30.66	30.89	30.69
B	Record Type - Surgical	2.28	2.27	2.31
C	Gender - Female	61.87	62.12	62.24
C	Gender - Male	38.13	37.88	37.76
C	Record Type - Ancillary	51.73	52.12	52.26
C	Record Type - Facility	0.62	0.62	0.63
C	Record Type - Management	24.77	24.87	24.67
C	Record Type - Pharmacy	21.17	20.73	20.77
C	Record Type - Surgical	1.71	1.65	1.66
D	Gender - Female	58.77	59.20	58.80
D	Gender - Male	41.23	40.80	41.20
D	Record Type - Ancillary	33.31	33.30	33.23
D	Record Type - Facility	0.39	0.40	0.39
D	Record Type - Management	22.37	22.36	22.31
D	Record Type - Pharmacy	41.52	41.56	41.67
D	Record Type - Surgical	2.41	2.39	2.39

Table 2: Percentage of Patients by Periods of Continuous Enrollment for Four Health Plans full and 10% data sets

Dataset	Continuous Enrollment	Full Dataset	10% Sample Dataset
A	$<$ 1 Year	30.59	27.38
A	1-2 Years	13.59	14.70
A	2-3 Years	11.33	13.44
A	3 - 4 Years	21.20	21.61
A	$>$ 4 Years	23.20	22.86
B	$<$ 1 Year	25.68	25.23
B	1-2 Years	5.59	6.78
B	2-3 Years	5.41	5.91
B	3 - 4 Years	63.33	62.09
B	$>$ 4 Years	0	0
C	$<$ 1 Year	21.61	28.79
C	1-2 Years	10.01	13.24
C	2-3 Years	5.76	10.33
C	3 - 4 Years	33.32	47.64
C	$>$ 4 Years	0	0
D	$<$ 1 Year	39.20	33.60
D	1-2 Years	14.75	16.15
D	2-3 Years	7.69	9.45
D	3 - 4 Years	31.45	30.45
D	$>$ 4 Years	10.71	10.34

The next step in evaluating the validity of the sample was to compare the production reports of the 10% sample to that of the full database. As you can see in the table below, the distribution of age by region was virtually identical for both the sample and the full database.

Table 3: Percentage of Unique Patients by Age Group (Year 1999)

Region	Data	Metric	0-9	10-19	20-29	30-39	40-49	50-59	60-64	65 and older
East	Base	Percent	17	15	12	17	17	13	4	6
East	Sample	Percent	17	15	12	17	17	13	4	6
Mid West	Base	Percent	18	16	13	16	17	12	4	6
Mid West	Sample	Percent	18	16	13	16	17	12	4	6
South	Base	Percent	14	12	12	16	17	13	5	11
South	Sample	Percent	14	12	12	16	17	14	5	11
West	Base	Percent	18	13	11	15	16	12	4	12
West	Sample	Percent	18	13	11	15	16	12	4	12

The next step in the process is to run a variety of the queries recently performed against the 10% and 1% sample and compare the results with those run against the full database. Detailed queries apply far more rigor to the testing than simple frequencies by one or two variables. Queries typically include a number of variables that result in relatively small counts. For example, a query may target any patient with diagnosis A and diagnosis B in 1999, with continuous enrollment 12 months before and 12 months after the triggering event and without a drug code A. We show now that the results of even complicated queries allow for accurate projection to the entire database.

Table 4: Results of Selected Queries on 1% sample database

Data	Query	Count of patients	Run time
Base	Diagnosis A	38238	
Sample	Diagnosis A	381	32 min
Base	Drug B	10828	
Sample	Drug B	111	25 min
Base	Drug C	94181	
Sample	Drug C	1001	38 min

Table 5: Hypertension Drug Groups, full and 10% sample database

Drug group	Full database	10% sample
Drug D	59126	5883
Drug E	66702	6589
Drug F	52038	5178
Drug G	46908	4700
Drug H	2187	219
Drug I	25056	2520
Drug J	2919	348
Drug K	18153	1706
Drug L	892	100
Drug M	1994	200

5 Bibliography

[1] B.Efron, R.J.Tibshirani,An Introduction to the Bootstrap, Monographs on Statistics and Applied Probability, No 57

[2] N.L.Johnson, F.C.Leone, Statistics and Experimental Design in Engineering and the Physical Science, J. Wiley and Sons, Inc. Vol.2, 1977

[3] W.G.Cohran, Sampling Techniques, J. Wiley and Sons,Inc. 1963.

[4] P.S. Levy, S.Lemeshow, Sampling of Populations, Methods and Applications, J. Wiley and Sons, Inc, 1999.

[5] P.Bickel, D.Freedman, Some asymptotic theory for the bootstrap, Annals of Statistics, 9, 1196-1217.

[6] J.Shao, Wo, C.F.J.Wu ,A general theory for jackknife variance estimation.Annals of Statistics, 17,1176-1197, 1989

[7] C.E.Lunneborg, Data Analysis by Resampling: Concepts and Applications, Duxbury Press,2000.

[8] E.I. Jilinskaia, T.J.Marx, S.J.Norton , Estimation of precision of Health-Care Cost analysis using multiple sampling methods, 2001, Proceedings of 4-th St. Petersburg Workshop on Simulation, 273 - 278.