

# Auditing The Auditors In Medicare And Medicaid

Aliza Rotenstein, Yeshiva University, USA  
Sheldon Epstein, Seton Hall University, USA  
Yonah Wilamowsky, Seton Hall University, USA

## ABSTRACT

*Baker Surgical Supplies, a small company, went bankrupt after it could not repay a significant overpayment charge demanded by Medicare based on a statistical extrapolation of claims of overpayment. The case centered on whether the extrapolation process was justifiable and whether it was properly implemented.*

*This paper provides a description of the extrapolation process used by Medicare and Medicaid and presents the data and sampling procedure offered by Medicare and the statistical arguments offered by Baker. The case demonstrates some potential misuses of statistics in the auditing process. In shedding light on this issue, the analysis in this paper could prove to be instrumental in prompting significant improvements to the auditing process of Medicare and Medicaid.*

**Keywords:** Auditing; Statistics; Sampling; Extrapolation

## BACKGROUND

In recent years, agents for the Centers for Medicare & Medicaid Services (CMS) have increasingly relied upon statistical extrapolation estimates when assessing claims of overpayments. Extrapolation is the use of statistical sampling to calculate and project overpayment amounts to be recovered by Medicare or Medicaid. Medicare's Program Integrity Manual (PIM) Section 3.10 "provides instructions for PSCs (Program Safeguard Contractors) and other Medicare contractors on the use of statistical sampling in their reviews to calculate and project (i.e., extrapolate) overpayment amounts to be recovered by recoupment, offset or otherwise." Section 3.10.1.4 continues to say that they "shall use statistical sampling when it has been determined that a sustained or high level of payment error exists, or where documented educational intervention has failed to correct the payment error."

In simple terms, when reviewing potential cases of overpayments to companies by Medicare and Medicaid, the sample size and selection methodology are first determined, and then records are randomly selected and requested from the provider. Statistically valid audit sampling methods must be used. These include simple random sampling, cluster sampling, stratified sampling, and systematic sampling. As a first step, an average of the overpayments within the sample is calculated. This average is then multiplied by the number of records in the overall population. Finally, a confidence interval is determined for the extrapolated amount so that there is a 90% probability that the total overpayment is at least a certain minimum. This amount is then demanded from the provider to be returned to Medicare or Medicaid.

Not infrequently, the sample size is quite small relative to the population size. For example, a sample size of 30 may be drawn from a universe of 1,000 to study the occurrences of overpayments. If the total overpayment in the sample is determined to be \$4,500, the average is calculated as \$150 per record (i.e., \$4,500/30). Multiplying the sample average by 1,000 then provides a point estimate for overpayment for the entire universe of 1,000. In this example, the point estimate is \$150,000 (i.e., \$150 x 1,000). Because this number is only an average overpayment, a Confidence Interval (C.I.), based on the mean and standard deviation of the sample, is calculated and the amount of repayment requested is the value at the lower end of the Interval. Thus, although only \$4,500 in erroneous claims is

actually demonstrated, if, in our example, the C.I. was calculated to be \$140,000 to \$160,000, Medicare would demand that \$140,000 be returned.

In 1990, the use of statistical sampling and extrapolation to calculate an overpayment was challenged in the case of *Chaves County Home Health Services, Inc. et al. v. Sullivan*, U.S. District of Columbia, U.S. District Court for the District of Columbia, (Feb. 12, 1990) Civil No. 86-2691. The Court ruled that statistical audits are permissible, and the ruling was upheld on appeal. As a result of the *Chaves* decision, demands by Medicare and Medicaid to return significant sums of money in overpayments, based on reviews of relatively small samples of sales, have sometimes threatened the viability of small and, at times, even large practices.

This paper discusses the statistical arguments presented by Baker Surgical Supplies, a small seller of durable medical equipment, in rebutting the demand of \$772,717.92 in overpayments calculated based on an extrapolation from a sample overcharge of \$130,325.30. Medicare’s demand of this large payment eventually put Baker Surgical Supplies out of business.

**THE CASE**

Baker Surgical was a durable medical supplies company specializing in wheelchairs, hospital beds, and comparable equipment. Table 1 lists requests that the CMS made to Baker to review a sample of records of past payments.

**Table 1: Requests by the CMS for Records**

<b>Date of Request</b>	<b>Number of Records</b>	<b>Dates Covered</b>
July 2002	5	1/1/2000 – 3/31/2002
May 2003	25	1/1/2000 – 3/31/2002
November 2003	31	4/1/2002 – 7/31/2003

Based on a review of these 61 records, on February 28, 2005, Baker was issued an overpayment notice with a demand for return payment of \$772,717.92. An appeal to the Administrative Law Judge in accordance with the procedures of Medicare was denied on May 8, 2006. A description of the statistical extrapolation and some of the arguments against it follows.

CMS claimed to have used a statistical package known as RAT-STATS to determine how many samples needed to be chosen, to select the samples to be taken, and to perform the extrapolation analysis. The projection was done in two parts: the first, combining the 5 and 25 records for the dates in January 2000 – March 2002, and the second, for the 31 records for the dates in April 2002 – July 2003. A stratified sampling technique was used with four strata for the first sample and three for the second.

According to PIM 3.10.4.1.3,

*Stratified sampling involves classifying the sampling units in the frame into non-overlapping groups, or strata. The stratification scheme should try to ensure that a sampling unit from a particular stratum is more likely to be similar in overpayment amount to others in its stratum than to sampling units in other strata. Although the amount of an overpayment cannot be known prior to review, it may be possible to stratify on an observable variable that is correlated with the overpayment amount of the sampling unit. ...an estimate is found for each stratum separately, and the weighted stratum estimates are added together to produce an overall point estimate.*

Section 3.10.7.1 further states that “the explanation of the sampling methodology that was followed shall include:

- a description of the universe, the frame, and the sample design;
- a definition of the sampling unit;

- the sample selection procedure followed;
- the numbers and definitions of the strata and size of the sample, including allocations, if stratified;
- the time period under review.

The description of the sampling procedure followed by CMS that was given to Baker listed the size of each sample, the size of each of the four strata, the overcharges found by CMS in each stratum, and the lower limit of the confidence interval that was extrapolated from a weighted average of the strata. In addition, the specific records chosen and the problem and amounts associated with those that were claimed to have overcharged were specified. A detailed report for the first group of 30 (based on the plaintiff’s report) is given in Table 2. The table shows that when the strata are combined, the point estimate of the total overpayment is \$549,472.08. Combining this with the overall standard error of the point estimate of \$27,494.52 (Total row, column (f)) results in a lower 90% confidence interval estimate of \$504,235.05 in overpayments for the entire universe of 507 reimbursements that took place in the January 1, 2000 – March 31, 2002 time period. Baker was requested to repay this amount.

A similar calculation was done for the second sample, covering April 1, 2002 through July 31, 2003. This is more briefly summarized in Table 3. When the strata are combined here, the point estimate of the total overpayment is \$303,595.47 and the overall standard error of the point estimate is \$21,346.96, yielding a lower 90% confidence interval estimate of \$268,482.87. Baker was requested to repay this amount as well. In total, based on 61 randomly chosen sales out of a total of 808, Baker was asked to return \$772,717.92.

**Table 2: Detailed Report on Overpayments for the First Sample (1/1/00-3/31/02)**

(a) Stratum	(b) Stratum Size	(c) Sample Size	(d) Mean Overpayment	(e) Point Estimate of Overpayment	(f) Standard Error	(g) Lower 90% C.I.	(h) Total Payment	(i) Net Proper Payment
1	144	1	\$218.44	\$31,455.36	\$0.00	\$31,455.36	\$34,266.65	\$2,811.29
2	120	4	\$384.78	\$46,173.00	\$11,635.36	\$18,790.78	\$71,794.97	\$53,004.19
3	157	7	\$1,130.03	\$177,414.26	\$4,582.95	\$168,508.76	\$168,345.67	(\$163.09)
4	86	18	\$3,423.60	\$294,429.46	\$24,494.58	\$251,818.53	\$329,543.10	\$77,724.57
Total	507	30		\$549,472.08	\$27,494.52	\$504,235.05	\$603,950.39	\$133,376.96

**Table 3: Detailed Report on Overpayments for the Second Sample (4/1/02-7/31/03)**

Stratum	Stratum Size	Sample Size	Mean Overpayment	Point Estimate of Overpayment	Standard Error	Lower 90% C.I.
1	116	4	360.36	41,801.76	5,866.21	27,996.44
2	125	7	855.96	106,994.82	9,960.10	87,640.55
3	60	20	2,579.98	154,798.89	17,946.49	123,767.05
Total	301	31		303,595.47	21,346.96	268,482.87

Note on Tables 2 & 3 that the numbers in the Total rows for Standard Error (column (f) in Table 2) and 90% Confidence Interval (column (g) in Table 2) are not the sums of the data from the individual strata above them. Rather, the total Standard Error is the pooled standard error taken from the individual strata. Similarly, the Confidence Interval is based on this pooled Standard Error.

**THE APPEAL**

Among other technical grounds, Baker appealed the case based on statistical arguments as well as claims that the audit did not follow the requirements in PIM. Baker’s statisticians argued that based on the evidence presented, the sample was not random, the sample could not be verified due to the lack of definition of the strata, and the statistical extrapolation methodology had serious flaws.

**1. The samples were not random.**

Other than stating that the sample was randomly drawn using RAT-STATS, no indication was given as to how this was done. Therefore, using the information they had, the plaintiff's statisticians compared the products and dates of the universe of records to those of the samples. Since the records of the universe satisfied a chi-squared test for uniformity over time, the samples were tested to see whether they satisfied this test as well. Three of the five originally chosen sales were from May 2001, and one each was from February 2000 and January 2002. When applied to this data, a chi-squared test using the Yates correction factor showed, with a p-value less than .01, that the data were not randomly drawn from the months in the entire period. Similar results were obtained for the remaining two samples as well as for the combined first sample of 30 records.

**2. The samples could not be verified due to the lack of definition of the strata.**

As mentioned previously, PIM requires an explanation of "the sample selection procedure followed, and the numbers and definitions of the strata and size of the sample, including allocations, if stratified." This is certainly reasonable if the sampling method is to be verified. Since no such definitions were supplied, there was no way to determine what the strata represented or whether the sampling procedure was adequate.

**3. The statistical extrapolation methodology had serious flaws.**

The stratified sampling procedure and analysis were deficient. While the totals of 30 and 31 items in each sample are fairly common as being sufficient, the individual sizes of the strata were not sufficient. Stratum 1 in Table 2 serves as a clear example of an extreme case. In this stratum, a sample size of one was chosen from a population of 144. While this is clearly insufficient on the face of it, there is an additional problem that negates the validity of the analysis. The standard error for this stratum is listed as zero (column (f)). A standard error of zero for the population is clearly meaningless. The result is a direct consequence of the fact that there is only one data item in the sample. When pooling the individual standard errors of the strata to get an overall standard error of \$27,494.52 for the total overpayment (Total row, column (f)), this zero, representing 144 out of 507 data items is included. This error in analysis will cause the overall standard error to be understated and thus increase the lower 90% estimate that is being demanded.

**4. The methodology led to 90% intervals that cannot be correct.**

For Stratum 3, the lower limit of the 90% C.I. for overpayment is given as \$168,508.76 (column (g)), i.e., we are 90% sure that the overpayment was at least this amount. This, however, is impossible, since the entire payment for this stratum was \$168,345.67 (column (h)). This impossible result provides further evidence of serious flaws in the sampling procedure.

Finally, in Section 3.11.1.2, PIM discusses the use of "probe samples." A probe sample is intended to be used as a preliminary study to determine whether further analysis is required. It is used to avoid the expense and hassle to the provider of more data retrieval than is necessary. Consistent with statistical norms, PIM requires that "contractors should include in the probe sample a random or stratified sample of generally 20 - 40 claims from that provider with dates of service from the period under review." Considering that the original request for five items was a probe sample, it is evident that in addition to the statistical flaws in the sampling procedure, regulatory requirements were not met either.

**RESULTS**

Upon weighing the evidence of the statisticians and accountants on both sides of the issue, the Administrative Law Judge (ALJ) ruled in favor of Medicare and denied all statistical and other claims of the plaintiff. Baker was not able to pay the amount of close to \$800,000 and was bankrupted. Similar cases have been similarly adjudged by ALJs in numerous instances. In a recent case, however, ALJ John Terepka issued a decision that has received a great deal of attention. In New York Department of Health, Appeal No. 08-3739, dated July 9, 2010, the ALJ ruled in favor of Rite Aid of New York. The decision refused to allow the Office of the Medicaid

Inspector General (OMIG) to extrapolate from a sample of audited pharmacy claims. In that case, the OMIG wanted to extrapolate from a sample to a population of around 135,000 and collect nearly \$300,000. The ALJ ruled that because the OMIG refused to disclose the details of the procedure by which it had selected the allegedly random sample, it could recoup only the amount of money actually found to be invalid in the sample, i.e., \$433.06. Along the lines of this ruling, our case similarly demonstrates some potential misuses of statistics in the auditing process. In shedding light on this issue, our analysis could prove to be instrumental in prompting significant improvements to the auditing process of Medicare and Medicaid.

#### **AUTHOR INFORMATION**

**Aliza Rotenstein** is an assistant professor of accounting at the Syms School of Business, Yeshiva University. She holds a PhD in business/accounting from the CUNY Graduate Center and teaches courses in financial and management accounting. Her main research interests include corporate governance, accounting restatements, and earnings management, and she is currently involved in research projects relating to accounting education and accounting for non-profit organizations. Her research has appeared in *Journal of Accounting Education*, *Journal of Applied Business Research*, and *Journal of Business Case Studies*. E-mail: arotenst@yu.edu.

**Sheldon Epstein** is a professor of computing and decision sciences at the Stillman School of Business, Seton Hall University. He holds a PhD in operations research from New York University and has done extensive business consulting. His research has appeared in: *Computers and Operations Research*, *Naval Research Logistics*, *Journal of the Operational Research Society*, *American Journal of Mathematical and Management Sciences*, *Opsearch*, *Annals of the Society of Logistics Engineers*, *Journal of Property Tax Assessment and Administration*, *Property Tax Journal*, *Location Sciences*, and *Interface*. E-mail: epsteish@shu.edu.

**Yonah Wilamowsky** is a professor and former chairperson of computing and decision sciences at the Stillman School of Business, Seton Hall University. He holds a PhD in operations research from New York University. His research interests center on applications of statistics and quantitative methods to the law, business processes, biblical studies, and higher education. Among other journals, his work has appeared in *Naval Research Logistics*, *Journal of the Operational Research Society*, *American Journal of Mathematical and Management Sciences*, *Tradition*, *Journal of College Teaching and Learning*, *Property Tax Journal*, *Location Sciences*, and *Computers and Operations Research*. E-mail: wilamoyo@shu.edu. Corresponding author.

**NOTES**