Compliance Risk: Pre- and Post-Audit Strategies

Frank D. Cohen, MPA, MBB
Senior Analyst
727.442.9117
frank@frankcohengroup.com
www.frankcohengroup.com

INTRODUCTION

Over the past several years, in an attempt to minimize the financial damage of the economic turndown, payers have increasingly turned to aggressive and robust audit events to minimize the amount that they pay providers for providing healthcare services to their subscribers. The federal government is no stranger to this process and with the introduction of the new health care reform legislation (AHCA), Medicare and Medicaid audits have increased in both their number and aggressiveness in order to fund the new program. This has resulted in an increase in both the number of providers being audited and the alleged overpayment amounts per audit. In all cases, the providers are treated as guilty until proven innocent as evidenced by the post-audit policies that require payment before appeal. In some extreme cases, such as with Eastern Carolina Internal Medicine (ECIM), the appeals process can take years and even when proven innocent, the cost to the provider can run into the millions of dollars. Referencing the above case, in 2008, AdvanceMed (an auditing entity that administers the CERT program) found that ECIM was overpaid by around a million dollars. In December of 2010, an administrative law judge ruled in favor of ECIM on over 90% of the reviewed claims, reducing the potential overpayment amount to just over $3,500. But by the time this happened, the practice had incurred hundreds of thousands of dollars in legal and financial costs and in the end, it took the intervention of a U.S. Senator to get CMS to pay the practice back moneys that they paid to CMS based on the flawed findings in the initial audit.

The above is just one example of many that occur far too often in our industry. Medical providers are at the mercy of audits conducted by unqualified individuals that are financially incentivized to find as many errors as possible and several studies and surveys show that somewhere between 35% and 80% of findings of overpayments are overturned in favor of the provider on appeal. It is critically important, then, that providers do two things; prepare in advance to assess, identify and mitigate your risk of an audit and know how to fight for your rights should the audit result in an unfavorable finding.

In this paper, we look at these two components in detail. The pre-audit risk assessment analyzes two areas; your risk for audit and the risk for an unfavorable finding. The post-audit review analyzes the statistical methodologies used when extrapolation is a part of the finding; a procedure that is increasing at an alarming rate.

WHAT IS AN AUDIT?

An audit, as presented here, is a review of medical claims submitted to a government or private payer. It normally consists of a review of the medical record to determine whether the procedure was actually performed, whether the documentation details the services billed and whether the documentation supports the medical necessity for the service or procedure for which the claim was submitted. Not all audits, however, involve the review of the chart. Automated audits are conducted using data found within the billing data, such as relationships between diagnostic and procedure codes, sex and/or age of the patient, the frequency with which a procedure or service is reported during a particular time frame or even the modifiers that may be associated to the line item in the claim.

There are several reasons that an audit event can be stimulated. Some are conducted as a random event. Others may be the result of a whistleblower. More often than not, however, we find that the audit event is precipitated as a result of benchmarking; that is, the payer is mining the billing data and finds anomalous occurrences for the tax ID associated with the claim (or claims). At times, it may be impossible to determine what triggered an audit, but nevertheless, you must always be prepared.

TYPES OF AUDITS

When factoring in private payer audits, the list of audit type can be quite long. Even just considering government payer audits, the list is, at the very least, quite imposing. They include:

- Recovery Audit Contractors (RAC)
- Zone Program Integrity Contractors (ZPIC)
- Medicaid Integrity Contractors (MIC)
- Medicare Administrative Carriers (MAC)
- Comprehensive Error Rate Testing program (CERT)
- Healthcare Fraud Prevention Enforcement Action Team (HEAT)
- Office of the Inspector General (OIG)
- Department of Justice (DOJ)

This is far from a comprehensive list and as stated above, excludes private payer audits. And while each of these entities is unique in their operational requirements, they also overlap with respect to jurisdiction. To get a better understanding of this, I have exemplified some of the audit entities below.

RECOVERY AUDIT CONTRACTOR (RAC)

First and foremost, it is important to know that RAC auditors are paid a commission based on how much they are able to recover from a provider. They are not, however, penalized if their findings are overturned in favor of the provider on appeal. As such, while RAC auditors are incentivized to find errors, they are not dis-incentivized to do so accurately and appropriately.

There are three types of RAC audits.

1. Automated Reviews use edit logic to process large numbers of claims without any review of the medical record
2. Semi-Automated Reviews begin as automated reviews but if a pattern of problems are discovered, quickly advance to complex reviews
3. Complex Reviews focus on claims identified as having a high probability of error and are manually reviewed

Not only are RACs involved in audits for Medicare Part A and Part B claims, they are now creating Medicaid RACs, which will focus on Medicaid claims and the potential for overpayments made to Medicaid providers.

It's important to note that, based on a recent survey of providers audited by RACs within the prior 12 months, 47.5% of claims determined to have been overpaid were reversed in favor of the provider on appeal. In essence, it is believed that RACs commit numerous errors in their audits and as such, providers should appeal every finding with which they disagree.

ZONE PROGRAM INTEGRITY CONTRACTORS (ZPIC

Unlike the RAC, the purpose of the ZPIC is to ferret out potential fraud and abuse and as such, ZPIC audits create a high degree of concern and anxiety amongst providers. ZPIC audits are most often referred by some other entity and in many cases, some behind-the-scenes reviews have already been conducted. ZPICs depend upon statistical sampling and extrapolation methods, which all but guarantee findings that reach significant financial levels. In many cases, pre-audit risk assessment can help to identify areas of risk within the organization and pinpoint foci for detailed internal review.

MEDICARE ADMINISTRATIVE CARRIER (MAC)

MAC audits are pre-payment medical reviews that are conducted to ensure that services provided to Medicare beneficiaries are both covered and medically necessary. It is a commonly held belief that MAC audits are often driven by the results of the Comprehensive Error Rate Testing (CERT) study, which identifies instances where carriers and Fiscal Intermediaries (FI) have paid practices in error. When errors are found, the practice is notified by the MAC and in most cases, the MAC requests the patient chart for a documentation review. According to CMS, MACs are supposed to be reporting their findings to the RACs. Therefore, if substantive errors are found by the MAC (prospective), this can stimulate a RAC audit (retrospective), which can significantly increase potential financial damages.

MEDICAID INTEGRITY CONTRACTORS (MIC)

MICs were established as part of the Medicaid Integrity Program (MIP) in 2005. The act requires CMS to contract with MICs to review and/or audit provider claims, identify improper and overpayments and provide education to providers, managed care entities and beneficiaries with respect to payment integrity and the quality of care provided. Interestingly, the MIC works at the federal level but on state-based issues. It is not, however, supposed to usurp an individual state's efforts at controlling fraud and abuse within their Medicaid program.

The MIC is supposed to ensure that paid claims were for covered services, coded and documented properly and paid in accordance with all current laws, policies and fee schedules. What is also interesting is that, even though this is a federally subsidized program, appeals are managed at the state level and the appeal process varies state-to-state. Appeals also vary based on organization type, i.e. hospitals, physicians, pharmacies, etc.

There are three types of contractors for this program:

1. Review MICs analyze claims data to identify payment vulnerabilities
2. Audit MICs conduct post-payment audits of documentation to identify overpayments
3. Education MICs educate the provider community as needed based on discovered issues

COMPREHENSIVE ERROR RATE TESTING (CERT)

The CERT program was established by CMS to monitor the integrity of payments made to providers by Medicare carriers and fiscal intermediaries. Included in this program is the Hospital Payment Monitoring Program (HPMP), which accounts for approximately 40% of the claims reviewed (the other 60% is under CERT). While HPMP focuses primarily on inpatient (or Part A) services, CERT focuses primarily on physician and outpatient (Part B) services.

The CERT study is supposedly based upon a random sample of claims from within the provider community. 'Supposedly' is used here because there are questions regarding the true statistical validity of its random sampling techniques. Then, for each record, CERT sends a letter to the provider requesting the medical record. If the practice does not respond after three attempts, the claim is recorded as paid in error, another practice that likely diminishes the validity of the CERT study. In fact, if a claim was found to have been under paid, it also is recorded as a payment error for purposes of the study.

For 2011, the study found that approximately 8.6% of all claims paid to all providers by Medicare were paid in error (95% CI 7.9% to 9.2%). For Part A, the proportion was 6.2% while for Part B, it was 9.2%. This accounted for $28.8 billion overall with $15.1 billion coming from Part A and $7.8 billion from Part B claims.

There are five error categories, as follows:

1. **No documentation**—the provider fails to respond to repeated attempts to obtain the medial records in support of the claim.
2. **Insufficient documentation—**the medical documentation submitted does not include pertinent patient facts (e.g. the patient's overall condition, diagnosis, and extent of services performed).
3. **Medically unnecessary service**—claim review staff identify enough documentation in the medical records submitted to make an informed decision that the services billed were ***not*** medically necessary based on Medicare coverage policies.
4. **Incorrect coding**—providers submit medical documentation that support a ***lower or higher*** code than the code submitted.
5. **Other**—Represents claims that do not fit into any of the other categories (e.g. service not rendered, duplicate payment error, not covered or unallowable service).

Not surprisingly, error categories for insufficient documentation and medically unnecessary service lead the pack with nearly 85% of all errors falling into this category.

THE PRE AUDIT RISK ANALYSIS

The pre-audit risk consists primarily of looking at the frequency with which both modifiers and procedure codes are reported for a practice. Additionally, we also consider the number of hours represented by the procedure utilization.

As discussed in the sections on specific audits, auditing agencies are more and more relying upon data mining and statistical analysis to ferret out aberrant claims and anomalous billing patterns. Some entities, such as ZPIC and OIG conduct detailed statistical analyses on claims to better understand the risk for fraudulent billing. This is evidenced by the new Fraud Prevention System (FPS) initiated by OIG in 2011. This system uses predictive analytics technology to review every Medicare fee-for-service claim prior to payment, for potential fraudulent and/or abusive practices.

Since June of 2011, this system has generated leads for 536 new investigations by CMS's program integrity contractors.

In conducting a pre-audit risk, the provider conducts their own utilization review to give them a better idea of what they look like to outside entities. The pre-audit risk assessment is designed to give the provider an idea of their compliance risk, provider performance.

DATA REQUIREMENTS

The analysis begins with a data set from the provider organization containing:

- Provider – lists each provider's ID, name and specialty
- Frequency – for each provider, the frequency with which s/he billed each procedure during the given time frame

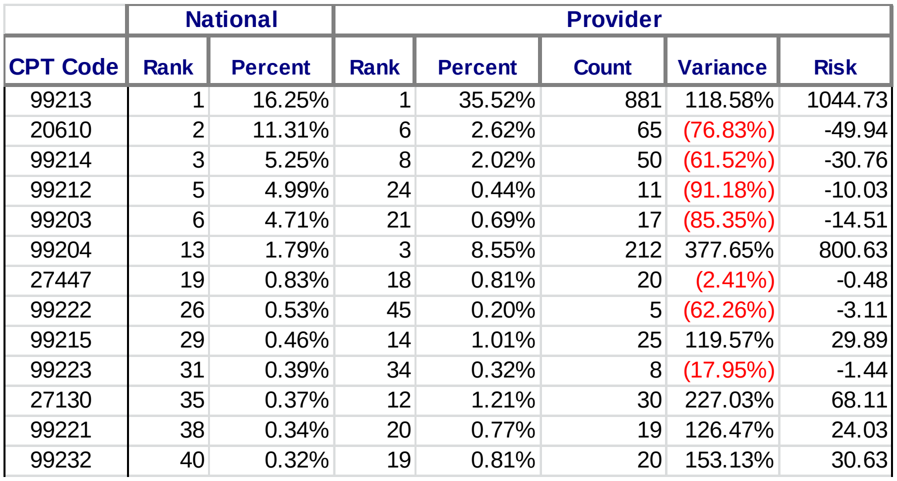
It is also necessary to obtain the control group, against which the provider's data will be compared. In most cases, this would be the CMS database (either Medicare or Medicaid) and can be obtained through either private companies or directly through CMS's website.

PROCEDURE CODE UTILIZATION

The idea here is to compare the utilization of procedure codes reported by the practice to that of some control group. As mentioned above, this most often takes the form of the CMS national claims database, also known as the Physician/Supplier Procedure Summary (P/SPS) Master File. This file contains 100% of claims submitted to CMS through the Part B claims system.

Critically, the top 10 to 25 procedure codes are compared against the control group. Statistical analyses, such as a chi-square goodness-of-fit test can be performed to determine whether the practice's variation from the control is statistically significant, however, in most cases, the practice conducts a 'smell test' by simply eyeballing the magnitude and direction of those variances. For example, let's say the most often reported procedure for a given specialty is code 12345, which accounts for 8% of the total count for all procedure codes and for the control group, this same procedure code for this specialty is ranked fourth with a proportion of 2%. In this case, we can see that, aside from the difference in ranking, which is likely to occur, the practice reports billing for this procedure code four times as often as their peer group. This variability is significant and would warrant a review of documentation to ensure that the procedure code is being documented appropriately and would withstand a medical necessity test.

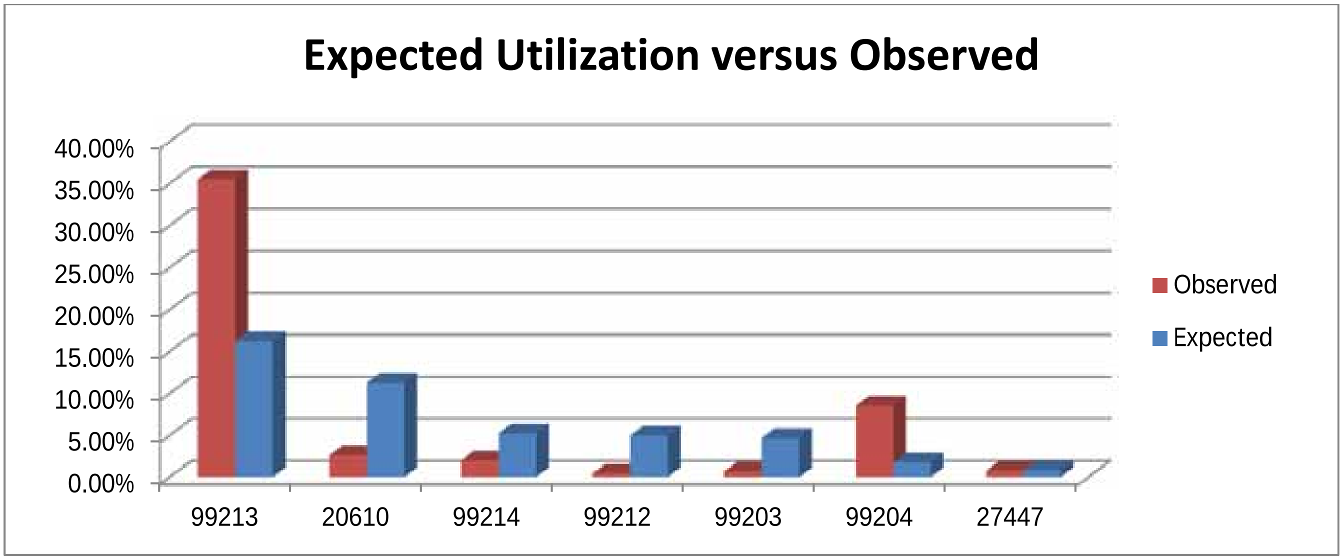
In some cases, one may wish to test both the direction and the magnitude of the variances and as mentioned above, a chi-square goodness-of-fit test is one way to do this. Few, however, will actually perform this test, rendering it a great idea but practically worthless. There are two other ways to get a handle on variance, one of which is through the use of graphics. Let's say you have the following table:



	National		Provider				
CPT Code	Rank	Percent	Rank	Percent	Count	Variance	Risk
99213	1	16.25%	1	35.52%	881	118.58%	1044.73
20610	2	11.31%	6	2.62%	65	(76.83%)	-49.94
99214	3	5.25%	8	2.02%	50	(61.52%)	-30.76
99212	5	4.99%	24	0.44%	11	(91.18%)	-10.03
99203	6	4.71%	21	0.69%	17	(85.35%)	-14.51
99204	13	1.79%	3	8.55%	212	377.65%	800.63
27447	19	0.83%	18	0.81%	20	(2.41%)	-0.48
99222	26	0.53%	45	0.20%	5	(62.26%)	-3.11
99215	29	0.46%	14	1.01%	25	119.57%	29.89
99223	31	0.39%	34	0.32%	8	(17.95%)	-1.44
27130	35	0.37%	12	1.21%	30	227.03%	68.11
99221	38	0.34%	20	0.77%	19	126.47%	24.03
99232	40	0.32%	19	0.81%	20	153.13%	30.63

One way to estimate the relational risk value for a procedure is to multiply the variance times the count. For example, for code 99213, the practice reports a variance 118.58% higher than their peer group. Multiply this by the number of times that procedure was reported and you get a value of 1,044.73. Now look at 27130 and notice that the variance is 227.03, higher than that for 99213. Multiply this by the count of 30 and you get a risk value of 68.11. You can also see this with 99204, where the risk value is a whopping 800.63. The point is this; the risk value is a function of both the variance and the frequency and suffice it to say that the higher the risk value, the greater the potential risk.

The second way is to look at the data graphically, as follows:



Note that this graph compares what we expected (based on the control group utilization) to what we observed (the actual proportion recorded by the provider). While not 'statistically valid', per se, it does give the analyst a pretty good idea of which codes, if any, may dominate the landscape.

TIME STUDIES

The RBRVS Update Committee (RUC) conducts a study that is updated every year and contains the number of minutes it takes to perform a given procedure code. In general, clinical experts from some 25 medical specialties are given a survey that contains 40 to 60 matching code pairs and asked to estimate the number of minutes it takes to perform each of these procedures or services. Time is broken into three components; pre-service, intra-service and post-service. These data are then used to extrapolate the analysis to over 7,000 procedure codes and in turn, used to create the work RVUs for those procedures.

Knowing this, then, it is a relatively simple exercise to estimate the number of assessed hours assigned to a given physician by getting the sum of the products of his/her minutes per procedure times the frequency for that procedure. OIG uses these estimates to identify physicians who report assessed times well in excess of what is believable. For OIG, fair market value calculations are conducted using 2,000 hours as a base. Recognizing that the RUC time study tends to be overestimated and that there are multiple work patterns among physicians, assessed times of over 5,000 hours significantly increase the risk to the physician for an audit or review. The reason? 5,000 hours would represent nearly 20 hours per day if the provider worked five days a week (no vacations) and around 14 hours a day if the provider worked 365 days a year.

MODIFIER UTILIZATION

Like procedure codes, the use of modifiers is subject to review based on utilization characteristics as well as qualitative rule sets. The main difference is the relationship the modifier maintains within the coding set. Procedure codes stand on their own; they are billed and reported independent of other codes. Modifiers, however, do not share this characteristic. A modifier is dependent upon the parent procedure code and therefore, utilization comparisons are not based on the total count of modifiers, per se, but rather the count of modifiers as a proportion to procedures.

Modifiers are also subject to certain qualitative rules. For example, some modifiers can only be billed with an E/M code, such as modifier 24 or 25. Others can never be billed with an E/M code, such as modifier 59. Some modifiers are subject to certain anatomical restrictions. For example, modifiers FA and F1 through F9 each identify a specific digit on either the left or right hand. E1 through E4 describe the upper or lower eyelid on either the left or right side.

The point is this; you should have a general understanding of modifier rules (or at least sets of modifiers) in order to make the analysis worthwhile. When I conduct an analysis, I restrict my review to those modifiers that I designate as high risk. These are modifiers that have either been reported as problematic by some agency, such as the OIG, or have repeatedly be the subject of audits and/or reviews in which I have been involved. In some cases, I may rely upon industry experts or litigation to identify a specific modifier as high risk. And when I create my utilization tables, I make sure to 'proportionalize' the relationship between the modifier and its parent code. For example, I calculate utilization for E/M only modifiers against only E/M codes and E/M-excluded modifiers against all but E/M codes. Conducting a qualitative modifier analysis is certainly important but a bit outside the scope of this presentation.

For my purposes, the following are high-risk modifiers:

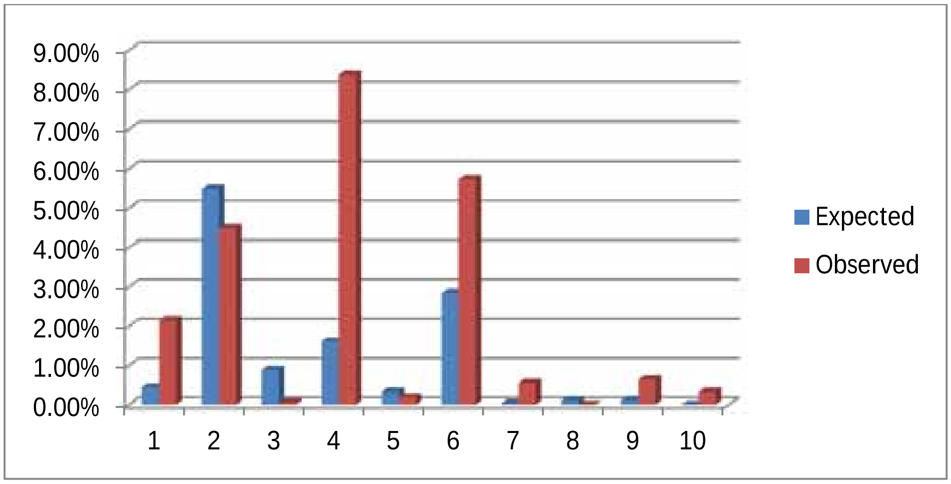
- -24 Use of E/M during Post-op Period
- -25 Separately Identifiable E/M Service
- -58 Staged/Related Procedure – Same Doc during post-op
- -59 Distinct Procedural Service (specific for CCI edits)
- -62 Two Surgeons
- -63 Procedure performed on infant < 4kg
- -76 Repeat procedure by same physician
- -78 Return to OR for related procedure during post-op

- -80 Assistant Surgeon
- -AS Assistant Surgeon – NP or PA
- -GE Performed by resident without physician supervision

In all other respects, modifier utilization analyses take on the same characteristics as procedure code utilization analyses, as seen in the table below.

Modifier	Expected	Provider Count	Observed	Variance	Risk
24 !	0.45%	35	2.15%	377.78%	132.22
25 !	5.49%	74	4.49%	(18.21%)	(13.48)
50 !	0.89%	1	0.05%	(94.38%)	(0.94)
51 !	1.61%	180	8.38%	420.50%	756.89
58 !	0.33%	4	0.19%	(42.42%)	(1.70)
59 !	2.84%	123	5.73%	101.76%	125.17
62 !	0.04%	12	0.56%	1300.00%	156.00
76 !	0.11%	0	0.00%	(100.00%)	-
78 !	0.12%	14	0.65%	441.67%	61.83
80 !	0.00%	7	0.33%	0.00%	-

As with procedure code utilization, we score the risk value based on the variance times the count. And as with the procedure risk assessment, we can also create a 'smell test' by looking at the relationship between what we expected (control group) and what we observe (actual for that provider).



THE POST-AUDIT ANALYSIS

Once an audit has been conducted, financial damage can be calculated in one of two ways:

1. Face value, assessing the potential overpayment for only those claims (or units) that were reviewed. For example, if 20 claims were pulled and documentation for those 20 claims reviewed, a face value finding could not exceed the actual paid value for those 20 claims.
2. Extrapolation can be used to estimate the value of the findings across the entire universe of claims from which the sample (of 20 in this example) is drawn.

Section 1842(a)(2)(6) of the Social Security Act requires the government to review, identify and/or deny inappropriate, medically unnecessary, excessive or routine services. Extrapolation techniques are used when the size of the universe of claims prohibits a complete review of every claim. In this case, a statistically valid random sample is drawn from that universe of claims in order to estimate potential payment error. In their "Standard of Work", CMS states that extrapolation may be used when there has been a determination that, within the universe of claims, there is a "sustained or high level of payment error" and again, this determination should be based upon a statistically valid random sample drawn from that universe.

The paragraph above says a lot; perhaps a lot more than most people initially see when they read it. First of all, it references a law that requires the government to conduct these audits. Second, it authorizes the use of a statistical methodology that surprisingly few people really understand. Thirdly, it creates a parameter for when extrapolation can be used and finally, it defines the criteria for invoking an extrapolated analysis. That's a lot to take in, especially considering that an entire volume can be (and has been) written to cover those four points.

Perhaps most alarming is the sentence that sets up the criteria for extrapolation. Again, it says, in part, that extrapolation can be used when there is a "sustained or high level of payment error" found within the sample. Alarming because, according to the statement of work, exactly what defines a "sustained or high level of payment error" cannot be challenged, either legally or administratively. I recently worked a case where 100 claims were reviewed and the auditor found four that were claimed as paid in error. Four out of 100 is not 4%; it's actually somewhere between 1.4% and 8.9% when you consider the concept of statistical error (these represent the range of the 90% confidence interval). The auditing entity decided to apply extrapolation and when the provider challenged with the reasoning that, at best, this represented less than a 2% error rate, he was told that he could not challenge what the auditor defined as "sustained or high level of payment error".

It's the third sentence that gives us hope; the part about being a statistically valid random sample drawn from the universe of claims. It is my experience that, in many cases, this is the best (and sometimes the only) way to challenge the application of extrapolation and again, in my personal experience, has a high rate of success. The reason is because of all of the audits in which I have been engaged, in only a handful did the auditing agency have a statistician involved or even someone who had the slightest idea of what constituted a statically valid random sample. So here goes!

SAMPLING

Probably the most accepted explanation of a random sample centers on the concept that, from within a given universe, every data point has an equal probability of being selected. Period. And in my experience, this is how most auditors approach this and also in my experience, they are usually wrong. This definition only holds when the universe from which the sample is to be drawn represents some semblance of homogeneity when regarding the characteristic of the included data points. In healthcare, this is hardly the most common finding.

Within the concept of random sampling is the technique of stratification. Stratification is a technique that separates the data points within a universe based on similar characteristics. For example, if we were looking to predict the winner of an election and the voting block covered multiple races, religions, nationalities, ages, etc., we would likely want to 'stratify' those blocks such that each were represented either proportionally or analyzed individually. For this example, you might want to score responses for African Americans v. Caucasian voters or those with a military background v. civilians or Jews v. Protestants. In the end, you can then extrapolate your findings back out to the entire population based on knowing the proportion of each of these stratified blocks within that universe.

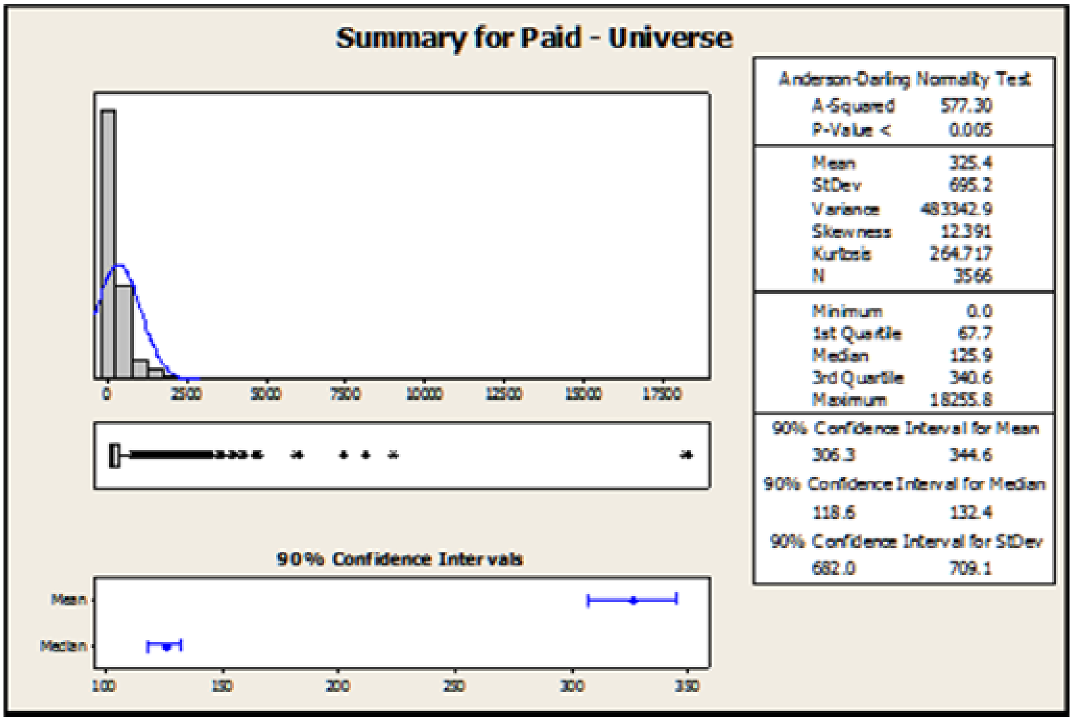
In healthcare, we face a similar dilemma, only the characteristics are not demographic or socioeconomic. They are based on code characteristics, payment levels, utilization, etc. For example, the American Medical Association (AMA) breaks CPT procedure codes into six major areas:

1. Anesthesia – procedure code range from 00000 – 09999
2. Surgery – procedure code range from 10000 – 69999
3. Radiology and imaging – procedure code range from 70000 – 79999
4. Lab and pathology – procedure code range from 80000 – 89999
5. Medicine – procedure code range from 90000 – 99200 AND 99500 – 99999
6. Evaluation and Management – procedure code range from 99201 – 99499

Why is this significant? Let's take a look at average payment amounts under the Medicare fee schedule. For the surgery procedures, for 2012, the median fee was $672.60. For E/M, for the same period, the median fee was $62.63; a ten-fold difference. Consider that, while E/M codes make up only 139 of the over 15,000 code groups found within the Physician Fee Schedule Database, they account for nearly a quarter of all claim lines. The point is this: in any given audit where extrapolation is being considered, E/M codes should ALWAYS be treated as an individual stratification and not combined in the sample with surgical codes.

Testing the average paid amount per claim is, perhaps, the most common of statistical tests conducted. Often, a two-sample t-test is conducted but even when that shows that the sample is statistically homogenous with the universe, it still does not mean that the sample is truly random. There are many other considerations and for the next few pages, I am going to address some of the issues and ways that a provider can move closer to ensuring that either the sample is, in fact, statistically valid or, consequently, move to have the extrapolation thrown out.

When looking at data for sampling, it is important to develop graphs for visual representation. Tables are fine and statistical tests are always important, but being able to visualize the data can often be the most important step in the process. Here's an example:



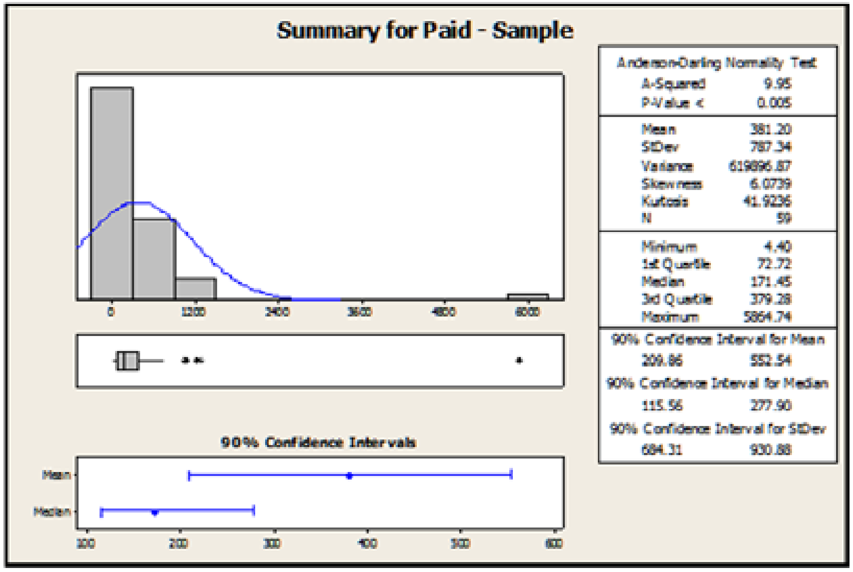
Summary for Paid - Universe

```
Anderson-Darling Normality Test
   A-Squared        577.30
   P-Value <         0.005

   Mean              325.4
   StDev             695.2
   Variance       483342.9
   Skewness         12.391
   Kurtosis        264.717
   N                  3566

   Minimum             0.0
   1st Quartile       67.7
   Median            125.9
   3rd Quartile      340.6
   Maximum         18255.8

90% Confidence Interval for Mean
   306.3             344.6
90% Confidence Interval for Median
   118.6             132.4
90% Confidence Interval for StDev
   682.0             709.1
```

This summary statistical analysis represents a universe of some 3,566 claims for a particular provider. Looking at the histogram alone tells an important story. For example, we see that the majority of the paid amounts hover around zero and in fact, there are many zero-paid claims within the universe. Think about the problem with this; if 30 claims are pulled at random and the average overpaid amount per claim is, say $25 then the extrapolation would (very simply stated) multiply that $25 times the universe of claims. In this above case, this would mean that, claims that were not paid at all would now be subject to a penalty of $25 each. In essence, the practice would be paying back money on claims that were never paid in the first place.

Second, notice the asterisks along the graph beneath the histogram. Each of these represent a statistical outlier. Again, this is totally unacceptable. Outliers should never be included in an extrapolation calculation. Why? Because they are, well, outliers. They possess some characteristic that moves them away from the requirements for a homogenous universe. Outliers should be assessed at face value only and not included in the sample.

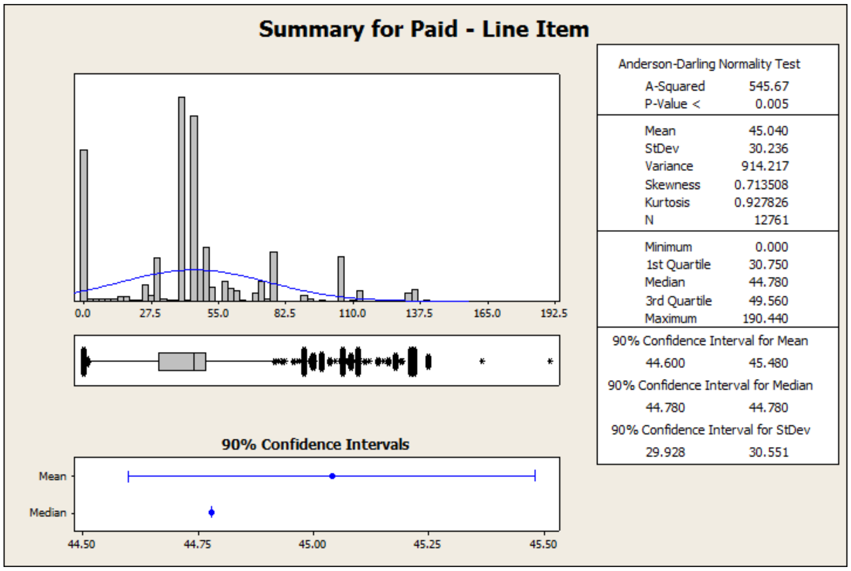
In fact, if we look at the histogram for the sample below, we can see that the sample contains at least four of the outlier claims.



```
Anderson-Darling Normality Test
  A-Squared          9.95
  P-Value <         0.005

  Mean             381.20
  StDev            787.34
  Variance      619896.87
  Skewness         6.0739
  Kurtosis        41.9236
  N                    59

  Minimum            4.40
  1st Quartile      72.72
  Median           171.45
  3rd Quartile     379.28
  Maximum         5864.74
90% Confidence Interval for Mean
  209.86           552.54
90% Confidence Interval for Median
  115.56           277.90
90% Confidence Interval for StDev
  684.31           930.88
```

Here's the problem with this. If that one outlier at the far right of the graph were found to have been paid in error, it could move the average overpaid amount per claim very far to the right. Let's say that it affects the total by $35, what would be the impact when extrapolated? Multiply that $35 times the number of claims in the universe (3,566) and you get an overestimate of nearly $125 thousand. That's a very big mistake and notice that it isn't in the provider' favor!

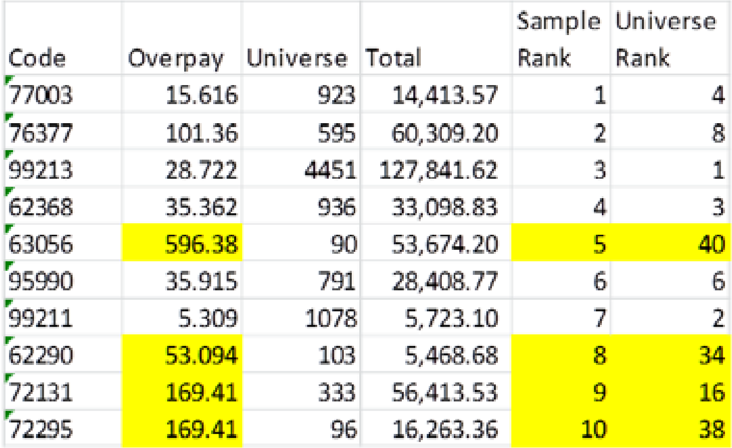
Here's an example of a case where stratification should have been applied.



```
Anderson-Darling Normality Test
  A-Squared        545.67
  P-Value <         0.005

  Mean             45.040
  StDev            30.236
  Variance        914.217
  Skewness       0.713508
  Kurtosis       0.927826
  N                 12761

  Minimum           0.000
  1st Quartile     30.750
  Median           44.780
  3rd Quartile     49.560
  Maximum         190.440
90% Confidence Interval for Mean
  44.600           45.480
90% Confidence Interval for Median
  44.780           44.780
90% Confidence Interval for StDev
  29.928           30.551
```

Note all the peaks in the histogram above. This is known as a multi-modal distribution. That is because of the numerous modes, or peaks, along the graph. Each peak likely represents its own separate distribution, which likely has its own set of characteristics that are separate from the others. Also note that zero-paid claims are included as well as outliers, both of which favor the auditor, not the provider.

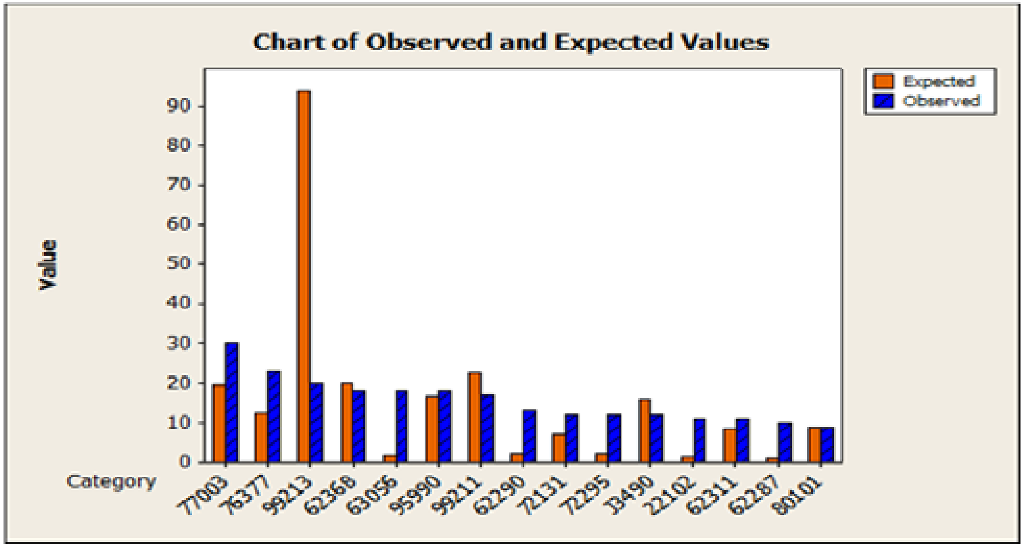
Another test, if you will, for validity has to do with the rank order of procedure codes within the sample compared to the rank order of codes within the universe. Below is an example:



Code	Overpay	Universe	Total	Sample Rank	Universe Rank
77003	15.616	923	14,413.57	1	4
76377	101.36	595	60,309.20	2	8
99213	28.722	4451	127,841.62	3	1
62368	35.362	936	33,098.83	4	3
63056	596.38	90	53,674.20	5	40
95990	35.915	791	28,408.77	6	6
99211	5.309	1078	5,723.10	7	2
62290	53.094	103	5,468.68	8	34
72131	169.41	333	56,413.53	9	16
72295	169.41	96	16,263.36	10	38

Note here that procedure code 99213 is ranked first within the universe, meaning that it was the most often reported procedure. In the sample, however, it is ranked as third. At the same time, procedure code 63056, which is ranked 40th in the universe is ranked number five within the sample. This is a big deal and quite often goes unnoticed. Note that the average overpaid amount for 99213 is $48.72. For 63056, it's $596.38. In an extrapolation example, then, you are going to see that 63056 has a huge influence over the total estimated overpayment when, in fact, it should probably not have even been included in the sample at all. At the same time, 99213, which has the third lowest overpaid amount will have a very small effect on the extrapolation when, in fact, it should be significantly higher.

When we conducted a chi-square goodness-of-fit test, it indicated independence between the sample and the universe, indicating that the sample was not a statistically valid representation of the universe.



Looking at the graph above, it is much easier to see the disparity between what we expected to see (based on the distribution of the universe) compared to what we did see (based on the distribution in the sample.

There are other issues to consider when looking for statistical validity within the sample and each should be looked at carefully. In practices with patients that run the gamut of severity, multi-stage cluster samples might be appropriate. Stratification should occur whenever there is evidence of independence between the sample and the universe. Zero-paid claims should not be included in either the sample or the universe when extrapolation occurs and outliers should only be assessed at face value. Any of these should stimulate an objection to the validity of the random sample, which, in turn, should be used to move to have the extrapolation set aside.

MEASURES OF CENTRAL TENDENCY

It is very common, within any data set, to try to find the approximate center of the data. In statistics, there are three main types of central measurements; the mean, the median and the mode. The mean (or the average) is perhaps the best known and the easiest to calculate. In calculating the average, one simply sums the values and then divides by the count of data points. The idea is that around half of the values are higher than the average while the other half is below the average. Hence why it is called the average. Averages also eliminate frequency bias, meaning that it is agnostic to how often a particular data point may have been reported. Sometimes, this is a benefit and other times it is not. In many cases, weighting the results for the frequency is very important and gives greater value to the results.

The other problem with averages is that they are only valid when the distribution of the data are either normal or symmetric; not a situation that we see very often in healthcare. Just take a look at the histograms displayed in the above examples. In my experience, appearances of normal distributions are~~is~~ a rare occurrence, indeed.

Rather, then, one should consider using the median rather than the mean. Here's the difference; the average measures the values while the median considers the position. Imaging you have a bunch of cells in a column of a spreadsheet. Each cell has a particular value. The mean looks at the value in each cell while the median looks only at the position of that cell in relation to all the other cells.

The procedure goes like this:

1. Place all the data points in a single column
2. Sort the data in ascending order (from low to high)
3. Count the total cells (number of records)
4. Pick the middle number (if an even number of cells and not a discrete data set, take the average of the two middle cells)
5. Open the cell and the value in that cell is the median

Medians are far less influenced by outliers and should always be considered when the distribution of the data are either non-normal or asymmetrical.

Picture the following data set:

31, 66, 71, 42, 91, 55, 65, 81, 99, 104, 19

The average (mean) is 62.5 and the median is 66.5. Pretty close, actually. But let's take the 105 and make it 1,040 instead. Now, the average is 156.10 while the median is 65.5, only one point difference from the original data set. In audit situations, we often see wide variation of paid claim amounts within both the universe and sample. In fact, the majority of data sets subject to audit are left bounded; meaning that they are limited on the lower end. For example, the least amount you can bill for a procedure is zero yet the maximum can be theoretically infinite. The most common distribution I encounter is a heavily left-skewed distribution, meaning that you see this long tail that extends way out to the right. In these types of distributions, the average is always higher than the median and that always (note the word always) benefits the auditor and not the practice.

The problem is that the program most often used by auditors (RAT-STATS) was designed by OIG for the purpose of selecting random samples and extrapolating results. Unfortunately, whoever designed it missed a very important point; that it is pretty much useless for extrapolation in the overwhelming majority of instances because the distributions are far from normal or symmetrical.

One should always push for the median as it tends to present the fairest representation of central tendency. And the simple fact is, in a normally distributed data set, the mean and the median are equal so there isn't any downside for using the median as a measurement of central tendency.

MEASUREMENTS OF VARIATION AND ERROR

In my opinion, a measurement of position, including that for central tendency, is pretty useless without including variation and error. This is particularly important when inferring the results of a test to a larger population, which is exactly what is going on with extrapolation. Variation measures the approximate distance between the individual data points and the central metric (mean or median).

VARIANCE

When we talk about the mean (or average), we use the standard deviation as the measurement for variance. Picture that you are standing in the middle of a room and there are other people there all around you. The standard deviation measures the average distance between you and all of the other folks. Some may be right next to you (within a foot or two) and other may be across the room (50 or 75 feet away). The way you would calculate this would be to take the distance from you to each person, square that value, add them all up and then take the square root of the total. In essence, we are measuring the dispersion of the data points; are they gathered closely around the central metric (small standard deviation) or scattered all over the place (larger standard deviation).

For the median, we use a different metric since the median doesn't measure values, it measures positions. In the prior example, we would take the distance from you to each of the other people in the room, place them in a spreadsheet and order them in ascending order. We would assign a name to each of the cells (the name of the person, perhaps). Then, we would get the total number of people in the spreadsheet (let's say 49 for ease of calculation) and count to the 25th position. Open the cell and the distance there would be the median distance. To calculate the variance, we use a completely different approach called the interquartile range (IQR). Here, we divide the data into quartiles (25th, 50th and 75th). The first quartile (25th percentile) identifies the point at which a quarter of the data points are lower and three quarters are higher. The third quartile (75th percentile) identifies the point at which three quarters of the data points are lower and one quarter of the data points are higher. The 50th percentile is the same as the median and we already know what that measures.

To calculate variance around the median, you simply subtract the first quartile from the third quartile. In effect, this identifies the range where 50% of the data points fall around the median. Using our example from above, let's say that the 25th percentile reported a distance of seven feet and the 75th percentile reported a distance of 38 feet. In that case, the IQR is 31 feet, meaning that approximately 50% of all the distances fall within that 31 foot range.

ERROR

In addition to variance, which gives us an idea as to just how dispersed the data points are around a central measurement, we also need to consider how accurate our estimates are, particularly when using a sample and then inferring the results of that sample to a larger database. There are, as you can imagine, a number of different statistical methods that are used to calculate error but for the purposes of the audit, the calculation is actually pretty simple. You take the standard deviation for the sample results and divide by the square root of the number of data points in the sample. For example, let's say that the RAC audits 30 claims. Of these, s/he finds that 12 of them have been paid in error and calculates a total overpayment amount of $1,232 resulting in an average overpayment per claim of $41.07 and a standard deviation of $32. You would take the $32 (standard deviation) and divide it by 5.477 (the square root of 30) to get a sample error of 5.84.

The next step is to convert this into a confidence interval. In general, the confidence interval is a range of values between which we have a certain amount of confidence, exists the true average (or median) for the universe from which the sample was drawn. In order to convert the sample error to the confidence interval, we need to multiply the sample error by a specific value and then subtract it from the mean (or median) to get the lower bound of the range and add it to the mean (or the median) to get the upper bound of the range.

Confidence intervals are represented by a percent value that defines our degree of confidence. For example, the FDA would often require a confidence interval of 95%, meaning that we are 95% confident (a weak way to explain this) that the true effect is somewhere between the bottom and the top of the range. For more critical applications, we may want to use a 99% or even a 99.9% confidence interval. In the audit world, at least for OIG and other government auditing entities, they rely on a 90% confidence interval. Note that the higher the confidence interval, the larger the range of values so a 90% confidence interval has a smaller interval (or range) than the 95% confidence interval.

The value that is used to convert the error to the confidence interval is driven by two things; the confidence interval itself (which is calculated as 1 – alpha, which is the risk of a false negative) and the degrees of freedom (or the sample size minus 1, sort of). To get the value, we use what are called the student's t tables. The t-value grows a bit smaller as the degrees of freedom (or sample size) get larger until, at infinity, for a 90% confidence interval, the value is 1.645. With degrees of freedom equal to 29 (sample size of 30 minus 1), the value is 1.699.

So, finally, we multiply the sample error of 5.84 times the t-score of 1.699 for a half-interval value of 9.93. Subtract this from the mean of $41.07 and you get a lower bound of $31.14. Add this to the mean and you get an upper bound of $51. Therefore, we could say that we are 90% confident that the true mean (or median) for the universe is somewhere between $31.14 and $51. Or, we would be nearly certain that if we were to analyze 100 samples of size 30 from this universe, that in at least 90 of them, the mean value per claim would be somewhere between $31.14 and $51.

Now, on to extrapolation!

EXTRAPOLATION

Let me start by commenting on the idea of extrapolation itself. While it has it proponents, there are many opponents within our industry. They claim that extrapolation is not a fair or reasonable method to determine the potential for improper payment and I respectfully disagree. I believe that their opinions are biased due to the fact that most, if not many, audits are conducted improperly, a high proportion of findings are in error, the samples are simply not random and the results of the extrapolations are, in fact, unfair to the provider. The truth is, if auditors were qualified and motivated to draw true statistically valid random samples and employ accepted statistical techniques, extrapolation would not only be fair but would significantly reduce the overall cost to both the auditors and the providers of having the audit conducted in the first place. It is simply unreasonable to conduct a retrospective audit of, say, 18,000 claims. This is unfair to the practice at least. Can you imagine the time and cost necessary to prepare, copy and submit every chart for every one of those 18,000 encounters? It may be a good strategy for a practice wanting to avert an extrapolation when they have this option as the cost to the auditing agency would be even more and they would likely not call the bluff. So, while I am not opposed to extrapolation as a technique, I am wary of the potential abuses it can invite.

As long as providers choose to be in a financial relationship with third party payers, then audits, recoupments and other recovery efforts will remain an active part of doing business in healthcare.

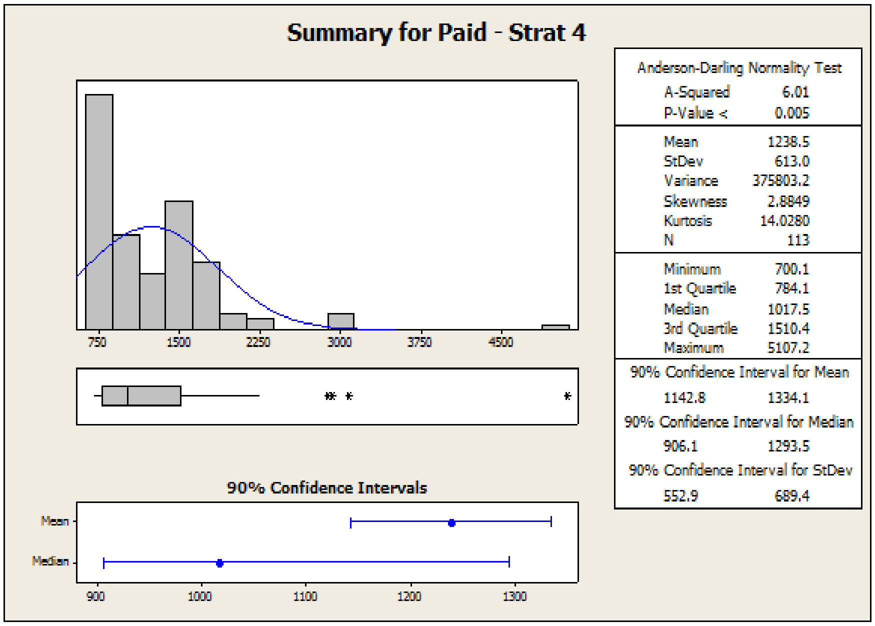
While there are some statistically accepted rules for performing extrapolations, different auditing entities may employ different methods. The two most often used depend upon either proportions of units found in error to the sample or point estimates of overpayment amounts for units found in error within the sample.

Let's start with the idea of using proportions. A couple of years ago, I was working on an audit conducted against a provider by Premera Blue Cross, a private payer in the Northwestern United States. In this one, Premera audited a sample consisting of 76 claims that were drawn from a universe of 3,489 claims. Ignoring that the sample was not random and that they committed lots of other quantitative and qualitative errors, their method for determining the extrapolation was totally different from any I had seen in the past. They determined that, of the 76 claims audited, 58.05% of them were paid in error. 58.05% of 76 equal 44.118, which didn't make sense since you can't have a fraction of a claim in error (these audits follow a discreet binomial rule which would have required a whole number). In any case, using an overly simplistic formula to calculate the proportion and then multiplying times the incorrect adjustment factor, they determined that, from this audit, the practice was paid in error on 49.26% of their claims. Multiply this percent times the total amount paid to the practice and they calculated the amount overpaid was nearly $600,000. When we asked them why they used this method instead of a standard accepted methodology, they responded that it always worked out to the provider's benefit. When I converted this analysis along with three others to a more standard methodology, in every case, the results favored the practice whereas the proportion method favored Premera. In this case, it reduced the estimated overpaid amount to just over $100,000. There are many statistical reasons why using a proportion approach such as this is simply wrong, from the heterogeneous nature of the claims to using the wrong normalizing method to an overly simplistic approach to calculating the proportion. Whatever the reason, the fact is it represents an unfair, statistically unsound practice and not only injures the provider but gives the idea of extrapolation a bad name. Remember, we judge any group of people by their least favorable members!

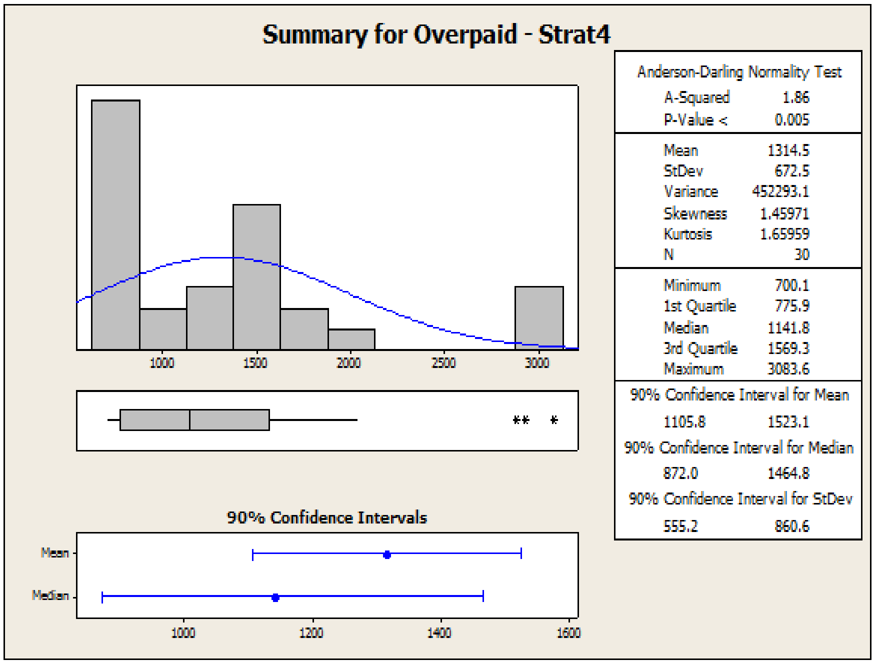
The preferred method and the one used by virtually all government auditing entities is to calculate the alleged overpaid amount per audit unit (i.e., claim line, claim, beneficiary, DRG, etc.) using some form of point estimate (i.e., mean or median), calculate the error and 90% confidence interval and extrapolation using the lower bound of the range, as discussed throughout this paper.

Needless to say, while the sampling, point estimate and error are critically important when determining the fairness and validity of an extrapolation analysis, the extrapolation methodology itself cannot be overlooked. And the best way to understand this is through case examples.

Even when a sample is stratified, it doesn't mean that it was stratified properly or that all other rules were followed. Recall that a sample can be stratified for a number of reasons but the most common is due to significant variation in the amount paid per claim. The histogram proved to be a great visual tool for determining the need for stratification due to this issue.



In this case, we plotted the data points for stratification 4 (paid claim amounts greater than $750) and then, below, for the sample.

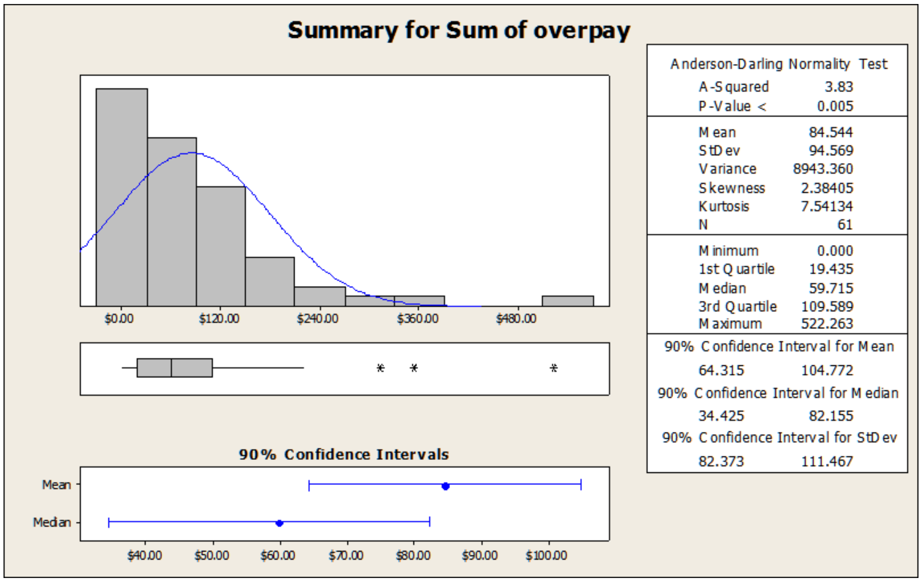


Note that the distributions are similar, which is good, but also note that nearly all (three out of four) of the claims identified as outliers in the universe ended up in the sample. And of these, all three were found to have been paid in error (or overpaid). This means that they were included in the calculation for average overpayment amount per claim, a mistake that not only biased against the practice, but resulted in a strange, yet not uncommon, paradox.

Whether we use the mean (average) or the median, look at the point estimates for the Summary for Overpaid graph above (the sample). The mean is $1,314.50 and the median is $1,141.80. Now compare that to the same statistics for the universe (Summary for Paid – Strat 4). The mean is $1,238.50 and the median is 1,017.50. In both cases, the estimated overpaid amount is greater than the paid amount, meaning that, if this extrapolation were allowed to proceed unchallenged, the practice would have to pay back more than they were paid. This is a great example of why outliers should not be included in the extrapolation analysis.

EXAMPLE 2

In the following example, there were several errors found that all biased the results against the providers. As always, we begin with a statistical overview of the data, as follows:

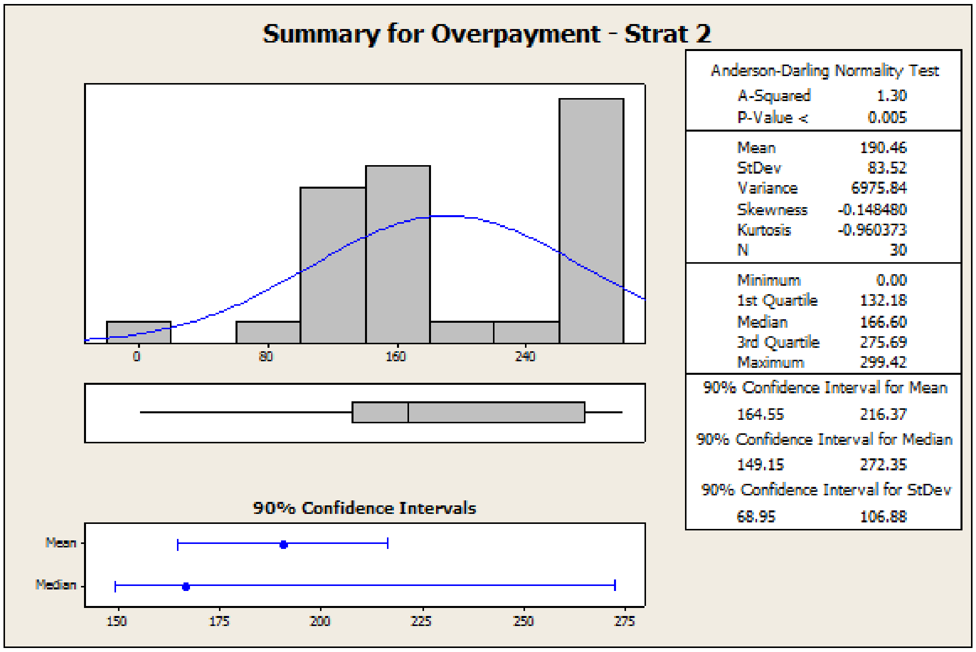


Anderson-Darling Normality Test	
A-Squared	3.83
P-Value <	0.005
Mean	84.544
StDev	94.569
Variance	8943.360
Skewness	2.38405
Kurtosis	7.54134
N	61
Minimum	0.000
1st Quartile	19.435
Median	59.715
3rd Quartile	109.589
Maximum	522.263
90% Confidence Interval for Mean	
64.315	104.772
90% Confidence Interval for Median	
34.425	82.155
90% Confidence Interval for StDev	
82.373	111.467

Here, we see that there are 61 claims that were audited. Of these, three were statistical outliers, which should have been addressed at face value and not included in the extrapolation. You can also see that the distribution follows the typical right-skewed distribution we most often see in these audits. For this audit, the sample was pulled from a universe of 12,011 claims so even a small error in the sample findings can result in a huge impact after extrapolation. Of particular concern, which builds on the last point, is the difference between the parametric (mean) and non-parametric (median) point estimates and confidence intervals. Note that the lower bound for the mean is $64.315 while the lower bound for the median is $34.425. When you extrapolate this to the universe of claims, the difference is $361,111 in favor of the practice ($772,487 using the average and $411,376 using the median). The main reason for this has to do with the way in which the point estimates respond to outliers with the mean values being affected far more significantly than the median values. This is another example of why the RAT-STATS program does not work for the majority of medical audits.

EXAMPLE 3

In this next example, we are looking at an audit that resulted in two stratifications. The first was for paid claim amounts between $80 and $300. The first problem is that there were zero-paid claims included in this stratification. Next, notice the atypical distribution of data as the data points are left-skewed. Also note that the histogram is multi-modal, which should have suggested that the stratification method was in error. When I reviewed the universe of the 4,293 claims, the results suggested that the second strata should have been between $80 and $200 and a third strata should have been proposed for paid claim amounts greater than $200.

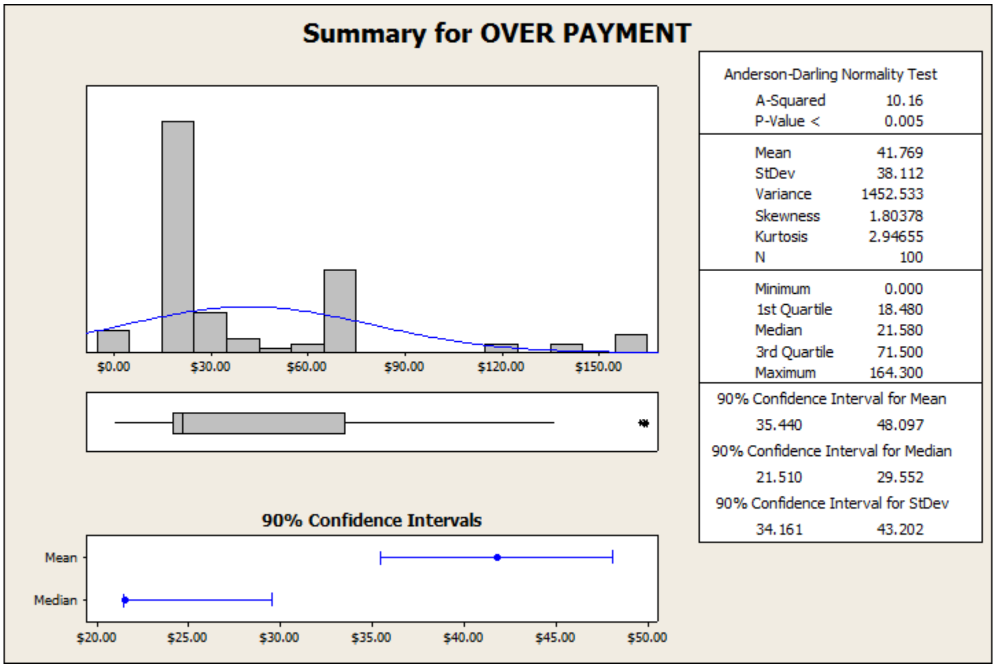


```
Anderson-Darling Normality Test
   A-Squared          1.30
   P-Value <         0.005

   Mean             190.46
   StDev             83.52
   Variance        6975.84
   Skewness      -0.148480
   Kurtosis      -0.960373
   N                    30

   Minimum            0.00
   1st Quartile     132.18
   Median           166.60
   3rd Quartile     275.69
   Maximum          299.42

90% Confidence Interval for Mean
   164.55           216.37
90% Confidence Interval for Median
   149.15           272.35
90% Confidence Interval for StDev
   68.95            106.88
```

Interestingly, even with a left-skewed asymmetric distribution, the median values were lower than the mean values. Here, the difference between the two lower bounds was $15.40 for a total impact of $66,112.20 in favor of the practice had the median value been used instead of the mean ($15.40 times the universe of 4,293 claims).

EXAMPLE 4

In this final example, we see a typical paid-claim distribution for a primary care practice. Note the relatively low financial value of the claim set. This sample of 100 claims was pulled from a universe of 10,256 claims.



Anderson-Darling Normality Test	
A-Squared	10.16
P-Value <	0.005
Mean	41.769
StDev	38.112
Variance	1452.533
Skewness	1.80378
Kurtosis	2.94655
N	100
Minimum	0.000
1st Quartile	18.480
Median	21.580
3rd Quartile	71.500
Maximum	164.300
90% Confidence Interval for Mean	
35.440	48.097
90% Confidence Interval for Median	
21.510	29.552
90% Confidence Interval for StDev	
34.161	43.202

Note the outliers found within the sample. These account for the significant difference between the mean (41.769) and the median (21.580) values and should have been excluded from the extrapolation calculation. Looking at the lower bound of the 90% confidence intervals, which again, are used in most cases to calculate the extrapolated damage estimate, we see a difference of $13.93. Multiply this times the universe of 10,256 claims and we calculate a total extrapolated error of $142,866.08 in favor of the auditing agency, not the provider.

TO APPEAL OR NOT TO APPEAL

Depending on what article, survey or study you read, the conclusion is that a very significant number of audit findings are overturned on appeal in favor of the provider. For physicians, this can range from around 35% to nearly 75%. For hospitals, the proportions are even higher. Imagine if you had a judge in your community whose decisions (findings) were overturned on appeal upwards of 75% of the time. Chances are, the only conclusion you could draw would be that the judge was likely not competent to make the decisions in the first place. My thesis is that this is the same situation with regard to audits. If over half of the audit findings are overturned in favor of the provider, then it would seem to me that the initial findings are simply in error a significant portion of the time. In a non-scientific survey that I conducted in 2012, I concluded that an appeal, on average, cost a practice $108 per claim. This is an average and includes the cost of going through all three levels of the appeal process. For RAC audits, in this same survey, respondents estimated that the average overpayment amount was $86. As such, in the best case, the practice loses around $22 for every successful appeal, a wholly unacceptable result.

It would seem that, under this scenario, that there would be some type of legislation that would require the auditors to reimburse the practice for the cost of the appeal beyond some acceptable proportion of error (I suggest anything beyond a 10% overturn rate). Without this type of protection, many practices choose not to appeal, which incentivizes the auditors to continue their pattern of (in my opinion) abuse.

WHY APPEAL?

The reasons that one should appeal should be obvious with one guiding principle; you disagree with the auditor's findings. And remember, there are two components of the appeal; qualitative and quantitative. The qualitative centers on a disagreement with the nature of the findings, whether it has to do with medical necessity, documentation, coverage rules, etc. Because of the elasticity inherent in the field of coding, we often see Clarke's fourth law of egodynamics invoked, which states "For every expert, there is an equal and opposite expert".

The fact is, all auditors are motivated, through different incentives, to find as many errors as possible and based on my many years of work in this area, far too many are not qualified to conduct statistical reviews. In general, if an auditor reviews 30 claims and finds 20 of those to be paid in error, an appeal would result in a reversal of those findings for anywhere between seven and 13 of those claims. This is why it is critical that you appeal every single claim where you don't agree with the auditor's findings.

HOW TO APPEAL

For any audit, there is supposed to be instructions on how to appeal included with the findings. And while it is not possible to provide an exhaustive set of instructions here on the appeal process, it is important to understand the three most common steps in the appeal process:

1. Redetermination, which involves a request for an independent (meaning other than the contractor staff involved in the original audit) source to review the findings.
2. Reconsideration, which involves having the findings reviewed by a Qualified Independent Contractor (QIC)
3. Administrative Law Judge (ALJ) hearing, which involves having the case presented in a trial-like setting before an ALJ. For Medicare, there has to be at least $130 in question.

For Medicare audits, there are two additional levels:

4. Appeals Council Review and
5. Judicial Review in U.S. District court. Note that, for 2012, the minimum amount in controversy must be at least $1,350.

CONCLUSION

With an increased concern over and focus on the cost of healthcare, every medical provider must be prepared to defend against an audit or attempt at recoupment. And with the proliferation of EMR/HER systems, provider data is more widely and easily accessible by the auditing agencies. In fact, our motto should be that of the U.S. Coast Guard; *semper paratus*, or Always Ready.

Conducting an *a priori* risk analysis, you should be able to assess two primary risks; the risk of being audited (how far you vary from your peers) and the risk of damage, which can only be truly assessed by an internal chart review. While you can't always do something to mitigate the risk of an audit, you can certainly do lots to mitigate the negative impact of an audit. Some practices avoid this up-front retroactive analysis out of fear of what they may find. While I am not a lawyer, my experience is that *ignorance of the law is not excuse*. In context, this means that, if you purposely avoid analyzing your data for fear of a negative outcome, the consequences can be worse than what an analysis might show. In most cases, if you do find overpayment issues on your own, self-disclosure is the best way to go. But if you don't plan to take action on negative findings, be prepared for the worst. So, while finding issues during your assessment may feel bad, not conducting an assessment may likely be worse. Your consequences, your decision.

If you are audited, you should ensure that the audit is as accurate and fair as possible by validating the sample, reviewing the audit results (qualitatively) and challenging those with which you disagree, verifying the point estimates and error rates and ensuring that the extrapolation methodology is reasonable and fair.

Whenever you disagree with an auditor's findings, whether at the qualitative or quantitative level, you should consider appealing the finding. To not do so not only costs you potential revenue, but it enforces the auditors' bad behaviors and makes it worse for your peers.

Audits are an inconvenience and a burden for medical providers but we should accept the fact that they are part of the cost of doing business in this industry. If you don't want to be audited, consider changing fields as, even with random audits, there isn't much you can do to avoid this nearly inevitable event. Unfortunately for the auditors, many believe that their motives are always bad, such as described prior regarding financial incentives. Alas, this is not always the case. Hanlon's Razor states the following: "*Never attribute to malice that which is adequately explained by stupidity"*. While paranoia can sometimes be a redeeming quality, we need to look past the conspiracy theories and try to focus on the reality of what happens during an audit. Purely sarcastically speaking, Hanlon's Razor can be restated as: *"Do not invoke conspiracy as explanation when ignorance and incompetence will suffice, as conspiracy implies intelligence."*