Basic Data Mining and Analysis: Data Integrity, Description, and Anomaly Detection

Data Integrity, Description, and Anomaly Detection

Health care enterprises process large volumes of data yet may have problems transforming data into actionable management information and business intelligence. Health care providers and other professionals also face a breathtaking array of data mining and analysis methods, tools, products, and services. As a result, while understanding that data mining and analysis help establish better controls, operationalize the "sentinel" effect, and demonstrate commitment to "doing the right thing right," it can be hard to know where to start the process.

Given program integrity's general objective to identify "what is not right," understanding what anomalies are and the basics of how to find them is a good place to start. In general, this ordered process should be a good first step.

1. Characterize the Question

Before considering "what data do I need," "where will I get it," and "what will I do with it," write down the question(s) you want to answer and then classify them by type. The classification scheme used by auditors, discussed below, can be quite helpful. It reveals how reviewers and auditors often structure their work and corresponds to the general flow of the data analysis process.

Audit questions fall into three general categories—descriptive, normative, or cause-and-effect:[1]

- Descriptive: Provides descriptive information about specific conditions of a program or activity;
- Normative: Compares an observed outcome to what is expected; and
- Cause-and-effect: Determines if observed conditions, events, or outcomes can be attributed to the operation of the program or activity.

Also:

- Descriptive analysis usually works with one variable at a time; that is, it is "univariate";
- Normative analysis usually works with two variables—the "norm" and "what is"—and is, therefore, usually "bivariate"; and
- Cause-and-effect analysis usually works with more than two variables and, as such, is "multivariate."

And, as we shall see, by classifying question types and the analyses to answer them, these three categories also help one choose which tool(s) to use. These three categories also lay the foundation for the general process of data analysis: describe it, then norm it, and then try to predict it. This process moves from relative simplicity toward greater complexity as shown in Figure 1.

Figure 1. General Data Analysis Model

Characterizing and documenting analytical questions help specify how and from where to collect the appropriate data, a topic not addressed in this document.

2. Control the Data

Good data analysis starts with good data. Even the best analytical tools cannot cure invalid or unreliable data. Data validity and reliability are bolstered by good information security (IS) including:

- Strong passwords;
- Good physical access controls, such as controlled and task- and staff-specific access to particular applications and administrative functions;
- Policies and procedures on tailgating, unrecognized persons, suspicious activity, workstation information control, and technology use while traveling and telecommuting;
- Means to identify and address cyber crimes such as identity theft, credit card abuse, spam, malware, hoaxes, cookies, Active X® applications, and phishing;
- Appropriate control over personal electronics and software at the workplace; and
- Firewalls, anti-virus and encryption software, and file back-up and retention.[2]

Health care data control includes protecting privacy using controls like:

- Accessing, using, and sharing information only on a need-to-know basis;
- Storing, transferring, and transmitting personal identifying information (PII) only by encrypted means;
- Confirming receipt of transmitted PII;
- Using social security numbers (SSNs) and healthcare identification numbers only at the time needed and never printing them in their entirety;
- Shredding, rather than merely discarding, documents containing PII; and
- Reporting and fully disposing of potential or actual privacy breaches and enacting appropriate disciplinary measures.[3]

It can be helpful to discuss your business, data and information needs, and technology infrastructure with your medical practice management software vendor to identify material threats, vulnerabilities, and risks.

3. Know Your System

The efficiency and effectiveness of data mining and analysis directly relate to the quality of the system used, electronic or otherwise. Finding, understanding, and controlling anomalies are the foundation of program integrity and require a measurement system that can:[4]

- Detect and show small changes (resolution);
- Respond to change in equal, constant, or appropriate ways (linearity);
- Show slow change over time (drift);
- Specify differences between high-side (upscale) and low-side (down-scale) values (hysteresis);
- Be consistent with like systems (difference among like gauges or like configurations); and
- Reveal variances in process, policy, practice, or people (difference among operations[tors]).

4. Know Your Limits

Before "crunching" data, determine the capacity and capability of the software used for mining and analysis to avoid "hitting the wall" in the middle of an analysis project. On the data capacity side, find out about the:

- Amount of computer memory required to run the software smoothly;
- Maximum number of available fields (columns) of data;

- Maximum number of available records (row) of data;
- Maximum number of characters available in a cell or mathematical formula; and
- Maximum number of files that can be open at the same time.

Regarding capability, learn about:

- Any special demands on, or requirements of, your operating system;
- What kinds of data (text and numbers) and file formats the software can handle;
- How easy or difficult it is to load, administer, change, and save files;
- Whether the tools needed are easy to get to without a lot of keystrokes or menu drill-downs;
- Whether basic arithmetic and graphs are easy to do;
- The types and numbers of formulas the software includes;
- How sampling is handled;
- How many levels of sorting can be done;
- Getting product updates and assistance; and
- The depth of detail and usefulness of the <Help> feature.

Be familiar with the existing practice management technology, especially before investing in new hardware or software. Find out if special or one-off analyses or reports are better created by customizing current software or exporting data to, and using, another application.

5. Know Your Data

Before mining or analyzing data, be sure to:

- Understand what each field (column) and record (row) contains and what they mean using a data dictionary, vetting to the data source, or other suitable means;
- Include a unique row counter, preferably in the far left-hand column;
- Save the original data set as a separate file and then work from a copy of that file;
- Formulate and document the question(s) you want to answer; and
- Delete irrelevant fields or records to make the file more manageable.

6. Assess Data Quality

What is "good" data? For auditors, data and the evidence they yield must be:[5]

- Sufficient—Is there enough data to persuade a knowledgeable person that the analysis and its results are reasonable?
- Relevant—Does the data have a logical relationship with, and importance to, the issue being addressed?
- Valid—Does the data give a meaningful, reasonable basis for measuring what is being evaluated?
- Reliable—Will the data and related analysis provide consistent results when information is measured or tested and are they verifiable or supported?

The quality of data mining and analysis can be enhanced by using multiple data analysis methods to help offset the weaknesses inherent in viewing or analyzing something in only one way. For example, interview evidence is more credible if supported by physical or documentary evidence.

Information from independent external sources is generally more reliable than from a single internal source.

7. Assess Data Integrity

The basic data integrity procedures below apply to any data set. Remember that data errors can exist but not matter, though they might point to control issues beyond the given analytical context. The importance of each data integrity issue hinges on several key questions:

- Does the error relate to this analysis or this question?
- If so, specifically how does it relate?
- Is this relationship material, that is, does it matter?
- Is this relationship significant? That is, does it matter enough to warrant seeking other data or changing or abandoning the analysis or the question?
- Does this type of error raise other important or future questions?

More specifically, inspect each column of data that matters to your analysis and look for:

- Blanks—Some fields, like claim numbers or patient identification numbers, should not be blank.
- Zeroes—Some zeroes are appropriate, others are not, particularly if a zero is a proxy for a blank.
- Error Values—Entries like #N/A, #REF!, #NUM!, and #NULL! can be inappropriate or indicate that data actually exists in the original source file but a calculation failed or data did not migrate.
- Unprintable Characters—On-screen data sometimes contain apostrophes, dashes, carats, or other characters that do not print but prevent data from functioning properly, especially when the data came from another (mainframe or legacy) computer application.
- Unnecessary Spaces—Spaces appearing before the first or after the last character in a cell can be residues of tabs, returns, or other commands that make even basic arithmetic impossible.
- Numbers Formatted as Text—Quantities imported as text often fail to calculate properly.
- Duplicates—Depending on the data extract, some fields and data values should not duplicate, such as claim numbers or dates.
- Edits—An overabundance of edits, corrections, and adjustments can highlight control or education issues or attempts to "cover one's tracks."
- Foreign Items—Data which are imported but were not actually requested or part of the data query are a possible indication of misunderstanding or a broader software failure.
- Unreasonable Values—This includes things like dates in the next century, ten-digit SSNs, two-digit Current Procedural Terminology (CPT) codes, numbers in names, a million-dollar copayment, etc.

Sorting a column ascending and then descending (or vice versa) often reveals such data errors.

8. Know the Data Type

Before choosing tools, set ground rules to help match data analysis method and purpose. Tool choice is driven by data type. Misinterpreting data type can result in inappropriate methods and create invalid, inaccurate, ineffective, or incorrect results.

Three general types of data exist:[6]

- Nominal data, or "attributes," use names, categories, or labels for qualitative values, such as gender, ethnicity, job title, etc.
- Interval variables, usually called just "variables," are true numbers, like dollar amounts or age. Nominal and ordinal data do not assert degree; for example, "one person is three times more male than another" or "person A said this training was five times more excellent than person B." Interval variables have meaningful value differences and allow statements about extent or degree.

- Ordinal data, “ranks,” are also categorical variables. The order of the categories has meaning, as in surveys using an ordinal scale ranging from “poor” to “excellent.” Such categories are often converted to numbers (4, 3, 2, and 1) for further analysis.

Program integrity data analysis usually involves only attributes that answer questions about compliance and control (was it done right, “yes” or “no”), and variables, that answer questions about volume and value (how many claims were done right), as distinguished in Table 1.

Table 1. How Attributes and Variables Differ

ATTRIBUTES	VARIABLES
Are qualitative.	Are quantitative.
Answer “yes/no” questions—“Are you male?”	Answer numeric questions—“How many males?”
Cannot assert degree—“I am twice as male.”	Can assert degree—“I am twice as old.”
Group information into classes.	Do not group information into classes
Can have only two values—“Yes” or “No.”	Can have an infinite number of possible values.

9. Describe the Data

There are two general types of descriptive statistics:[7]

- Measures of central tendency—Where data tends to fall.
- Measures of spread—How spread out or concentrated the data are.

Three central tendency measures are common, and each works best with a given data type:

- Mean (average value)—Best measure of a variable (quantity); for example, how much did the patient pay for each visit last year on average?
- Mode (most frequent value)—Best measure of an attribute (quality); for example, is the patient on Medicaid?
- Median (middle value)—Best measure of a rank; for example, does the patient rate services as excellent, good, fair, or poor?

Common measures of data spread include:[8, 9]

- Range—Difference between largest and smallest values.
- Interquartile range—Difference between the 75th and 25th percentile.
- Standard deviation—Square root of squared average differences between each datum and the mean —34 percent of the mean on the bell curve.
- Skew—Measures whether data are symmetrical to the left and right of center—zero on the bell curve.
- Kurtosis—Measures whether the data are peaked or flat relative to a normal distribution—zero on the bell curve.

On the bell curve or normal distribution, “the most important and most frequently used distribution in statistics,”[10] the more different the mean, median, and mode become, the higher the spread values get, the farther the skew and kurtosis get away from zero, and the less normal the data set. Skew and kurtosis are considered “large” if they are at or above 3 and 4, respectively.

The less normal the data set, the more possible, though not certain, non-compliance or improper payment may be. Data abnormalities, often called "anomalies" or "outliers," are of great interest. The descriptive statistics in Figure 2 below are from an abnormal data distribution. The mean, median, and mode are far apart, the standard deviation is large relative to the mean, and the skew and kurtosis are materially different from zero.

Figure 2. Descriptive Statistics Example

ANNUAL DOLLARS BILLED PER PATIENT	
Mean	$939.88
Median	$161.33
Mode	$112.20
Standard Deviation	$1,987.98
Kurtosis	90.52
Skewness	6.98
Range	$38,595.60
Minimum	$6.06
Maximum	$38,601.66
Interquartile Range	$856.66
Sum	$4,699,416.34
Count	5,000

10. Look for Qualitative Anomalies

The characteristics or qualities of data, not just their values, can be anomalous, as with an excluded person or entity on such lists as:

- U.S. Department of Health and Human Services, Office of Inspector General (HHS–OIG) List of Excluded Individuals/Entities excluded from Federal health care programs at http://oig.hhs.gov/exclusions/ on the Internet;
- Federal System for Awards Management list of debarments from all Federal programs at https://www.sam.gov/portal/SAM/#1 on the Internet; and
- General Services Administration's State-level suspensions and debarments at https://www.gsaig.gov/index.cfm/suspension-and-debarment-listed-by-state on the Internet.

Anomalies can also relate to demographic and geographic characteristics such as the following items of possible interest when looking for inappropriate relationships in medical claims data:

- Unduplicated beneficiary numbers per beneficiary name and vice versa;
- Dates of birth and first date of service per beneficiary name and number;

- Addresses per beneficiary name and beneficiary number and vice versa;
- Claims, codes, and diagnosis-related groups per beneficiary name and number; and,
- Total billed, paid, third-party liability and collection, copayment, and crossover payments per beneficiary name and beneficiary number.

11. Look for Quantitative Anomalies

Quantitative "outliers have extremely large or small values that place them on the outer reaches of the distribution."[11] Some common ways to identify quantitative outliers are quantities like claim counts, date-of-service counts, payment amounts, etc., that lie beyond:

- 3 standard deviations above (or below) the mean;[12]
- 3.5 standard deviations above (or below) the median;[13]
- The "inner fences;"
- The "outer fences;"[14]
- A certain percentage above (or below) the mean or median; and
- A certain percentage increase (or decrease) in a given variable.[15]

Besides basic value and volume of claims, clients, details, and dollars, other variables to consider in health care data when scanning for outliers with these criteria are:

- Average number of claims, line-item details, dates of service, and dollars per beneficiary;
- Average number of line-item details, dollars, and beneficiaries per claim;
- Average dollar value of a beneficiary, claim, and line-item detail; and
- Average beneficiaries, claims, line-item details, and dollars per unit time.

12. Address the Anomalies

In dealing with outliers:

- Investigate them carefully for information on the process or the data themselves;
- Try to determine their source(s) and reason(s) for their appearance; and
- Find out if they are control/compliance errors or simply "bad data."[16]

Given that outliers, especially from measurement error, can make data mining and analysis misleading, they should be addressed early in the process. Possibilities include:[17]

- Using metrics that are relatively insensitive to outliers, such as the median (rather than the mean);
- Eliminating them from the analysis; and
- Treating them as a separate group, especially if the analysis will drive decision-making, since outliers can skew the results of sampling or other analytical assertions.

13. Do the Graph

Graphs are the basic form of data visualization. A good summary discussion is available from the U. S. Geological Survey at http://pubs.usgs.gov/twri/twri4a3/pdf/chapter2new.pdf on the Internet. Though numerous types are available in today's software, most basic analyses can get by with just two—a pie chart (refer to Figure 3) for attributes and ranks and a histogram (refer to Figure 4) for variables.[18] The histogram is especially useful because it shows:

- Central tendency;
- Spread;
- Skew;
- Possible outliers; and
- Presence of multiple modes (another possible form of abnormality).

Figure 3. Pie Chart Example

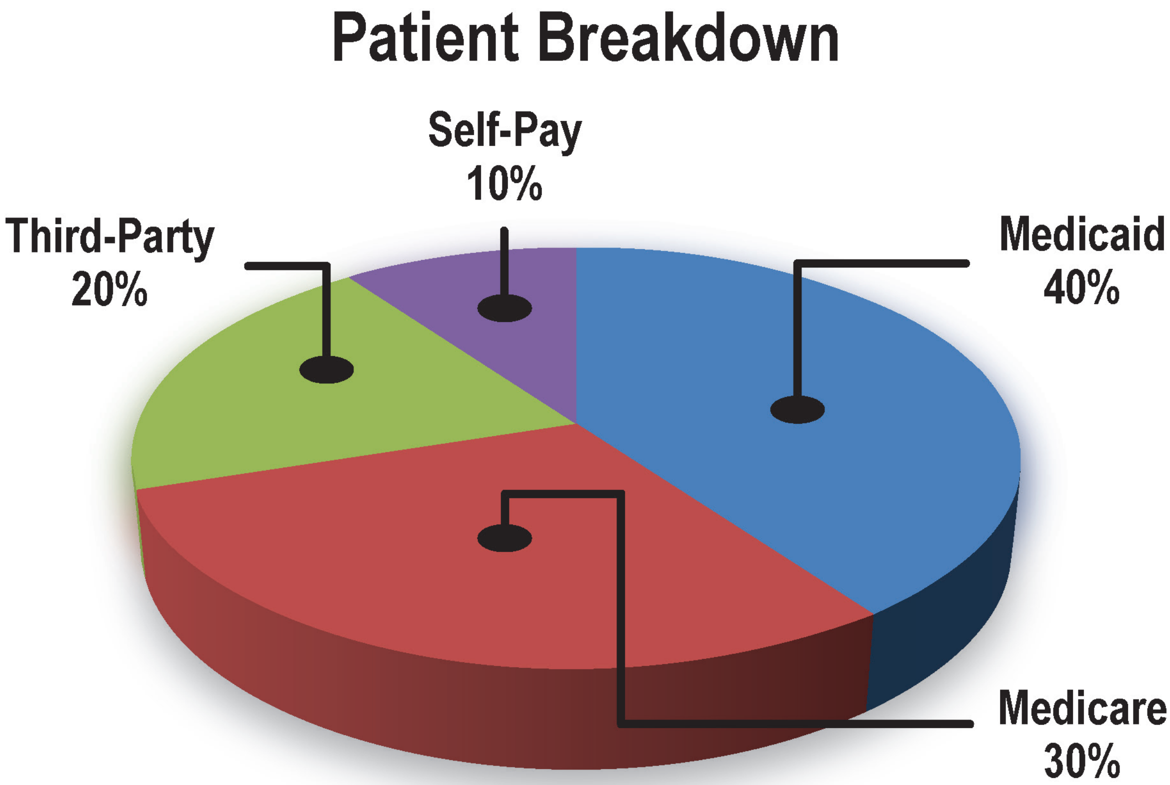
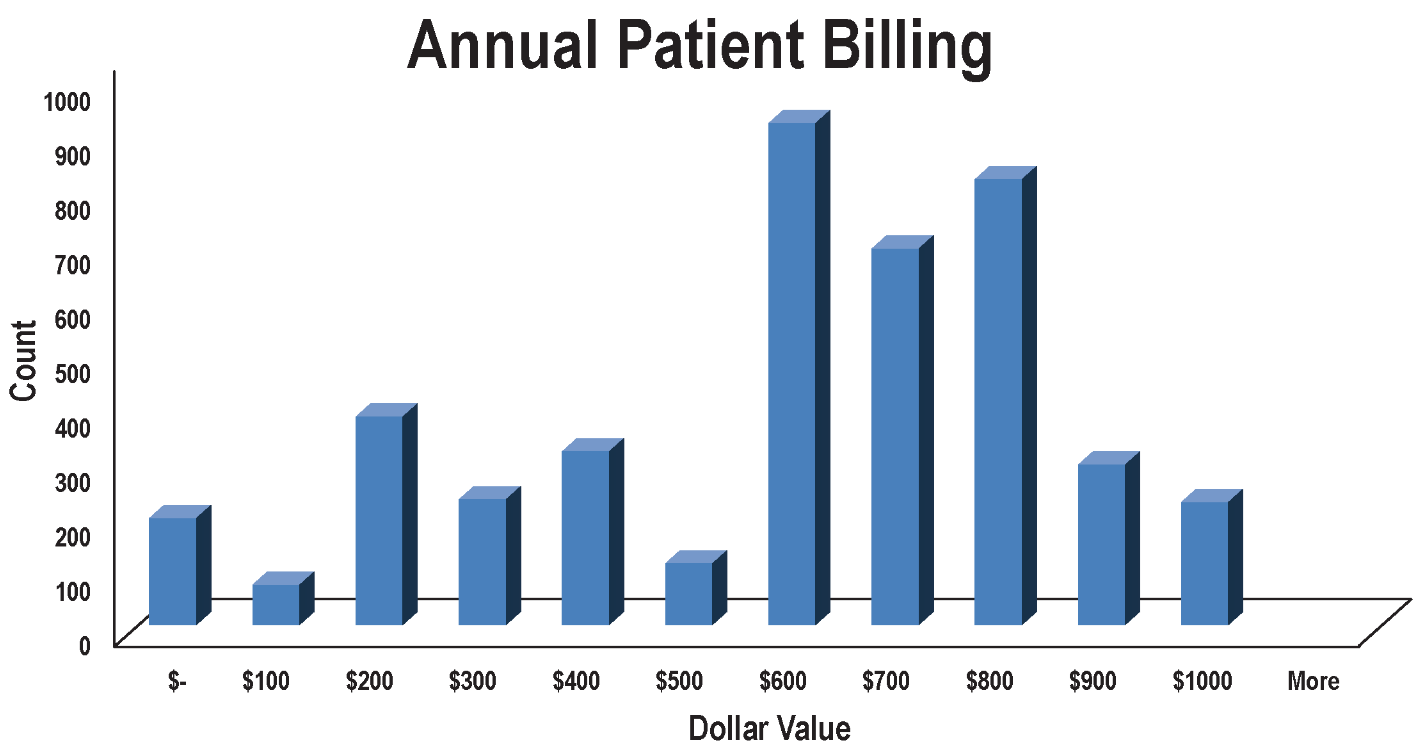


Figure 4. Histogram Example



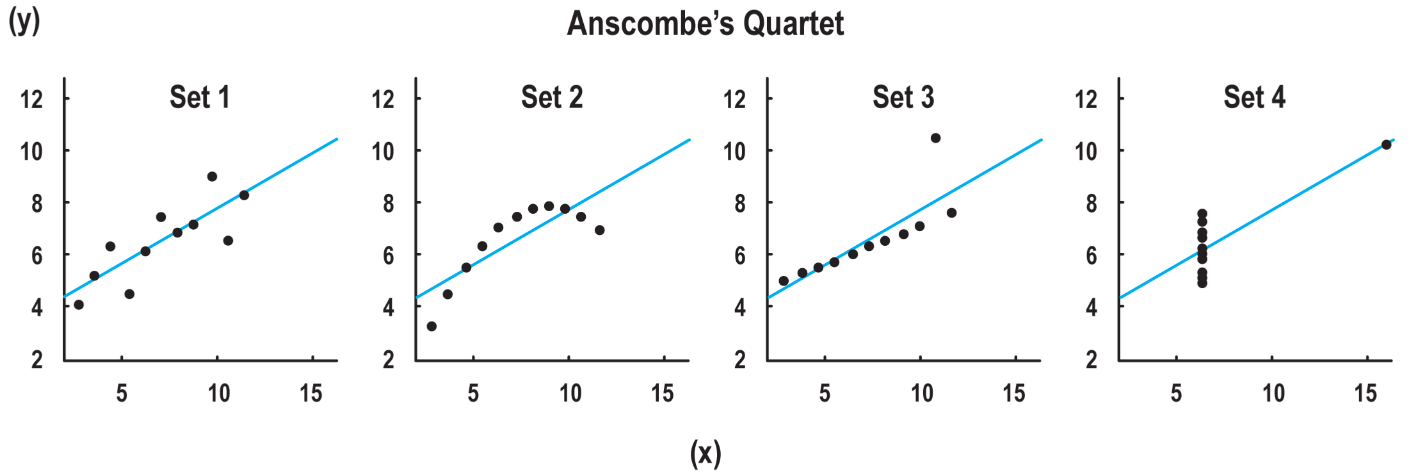
A famous demonstration of the pivotal importance of graphs is “Anscombe’s Quartet.”[19] The four data sets shown in Figure 5 are nearly identical.

Figure 5. Anscombe’s Data

ANSCOMBE’S QUARTET—DESCRIPTIVE STATISTICS								
Statistic	**X - 1**	**Y - 1**	**X - 2**	**Y - 2**	**X - 3**	**Y - 3**	**X - 4**	**Y - 4**
Count	11	11	11	11	11	11	11	11
Sum	99.00	82.51	99.00	82.51	99.00	82.50	99.00	82.51
Mean	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Median	9.00	7.58	9.00	8.14	9.00	7.11	8.00	7.04
Standard Deviation	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
Variance	11.00	4.13	11.00	4.13	11.00	4.12	11.00	4.12
Standard Error	1.00	0.61	1.00	0.61	1.00	0.61	1.00	0.61
Range	13.50	5.08	13.50	4.70	13.50	5.08	13.44	4.57
Sum	0.33	0.22	0.33	0.22	0.33	0.22	0.33	0.22

Given that all X-values are essentially equal, as are all the Y-values, we expect the scatter-plot graphs to resemble one another. However, this is clearly not the case, as shown in Figure 6.

Figure 6. Anscombe’s Graphs



Both descriptive statistics and a graph should be used, and they should “say” basically the same thing, especially given that graphs are easily manipulated to one’s end.[20] Remember, too, that descriptive statistics are “descriptive.” They say little, if anything, about whether norms are met or about cause and effect. Making big changes based on mere data description is often unwise. Additional measurement, mining, and analysis are usually needed.

14. Explore Data Relationships

After separately analyzing each important attribute, rank, or variable (univariate analysis), we can explore relationships and patterns between them starting with two variables (bivariate analysis). Again, the method used depends on the type of data. Correlation matrices are a good place to start. They reveal how much and in which direction one value changes when another value changes. The values vary between -1 and +1, with zero (0) indicating no relationship.[21]

The objective is to consider the strength and direction of the correlation value in each cell in the matrix and consider:

- Does a relationship (not) exist where one should (not)?
- Are the strength and direction of the relationship reasonable?

Most software populates only half of the matrix to avoid redundancy, and the values of "1" are expected given that a variable correlates perfectly with itself, as shown in Figure 7.

Figure 7. Correlation Matrix Example

VARIABLE	Payment Count	Payment Date	Check Number	Paid Amount
Payment Count	1			
Payment Date	0.0187	1		
Check Number	0.0186	0.8999	1	
Paid Amount	0.0029	0.0131	0.0130	1

The example causes concern about the imperfect relationship between Check Number and Payment Date. Both quantities should be perfectly sequential but do not show a correlation materially equal to 1. This suggests a possibly inappropriate number of voided, missing, or unprocessed checks.

Correlation is not causation. Consider the absurdity of a matrix of the same numbers that asserts a strong relationship between "Ice Cream Cones Eaten" and "Flights to the Moon," as shown in Figure 8.

Figure 8. Correlation Is Not Causation

VARIABLE	Patient Height	Flights to the Moon	Ice Cream Cones Eaten	Fish Caught Last Year
Patient Height	1			
Flights to the Moon	0.0187	1		
Ice Cream Cones Eaten	0.0186	0.8999	1	
Fish Caught Last Year	0.0029	0.0131	0.0130	1

15. Explore Comparisons

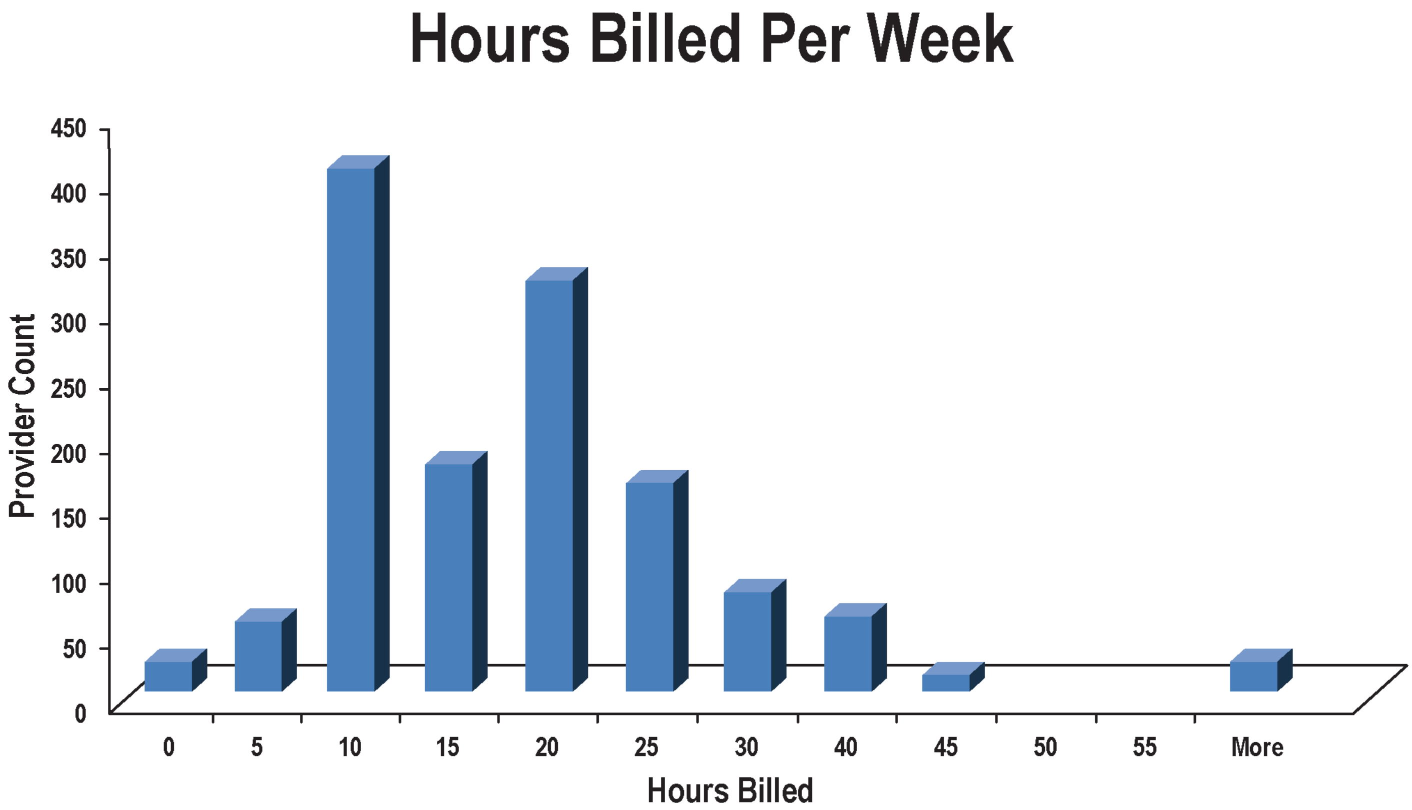
To know about an attribute, look at percentages. For example, look at the percentage of claims that are or are not processed within a given time period in a multi-facility practice. Figure 9 suggests that the Hilltop Clinic tends to bill late while Valley View does better.

Figure 9. Attribute Analysis Table

Billing Clinic	Payment Aging			Totals
	0 - 30 Days	30 - 60 Days	Over 60 Days	
Hilltop	23%	31%	47%	**100%**
Woodlands	30%	34%	36%	**100%**
Valley View	41%	31%	28%	**100%**

With variable data, such as dollar amounts, in addition to the quantitative outlier anomaly metrics discussed earlier, histograms are valuable. For example, by analyzing the weekly hours billed by clinical psychologists, the bar at the far right of the graph may be anomalous, as shown in Figure 10.

Figure 10. Variable Analysis Graph



16. Explore Patterns and Trends

A common method for looking at patterns with attribute (nominal) or rank (ordinal) data is the cross-tabular display,[22] or simply "cross-tab." The numbers in the cells of the table indicate the number of customers responding to each possible combination of attitude and wait time, as shown in Figure 11.

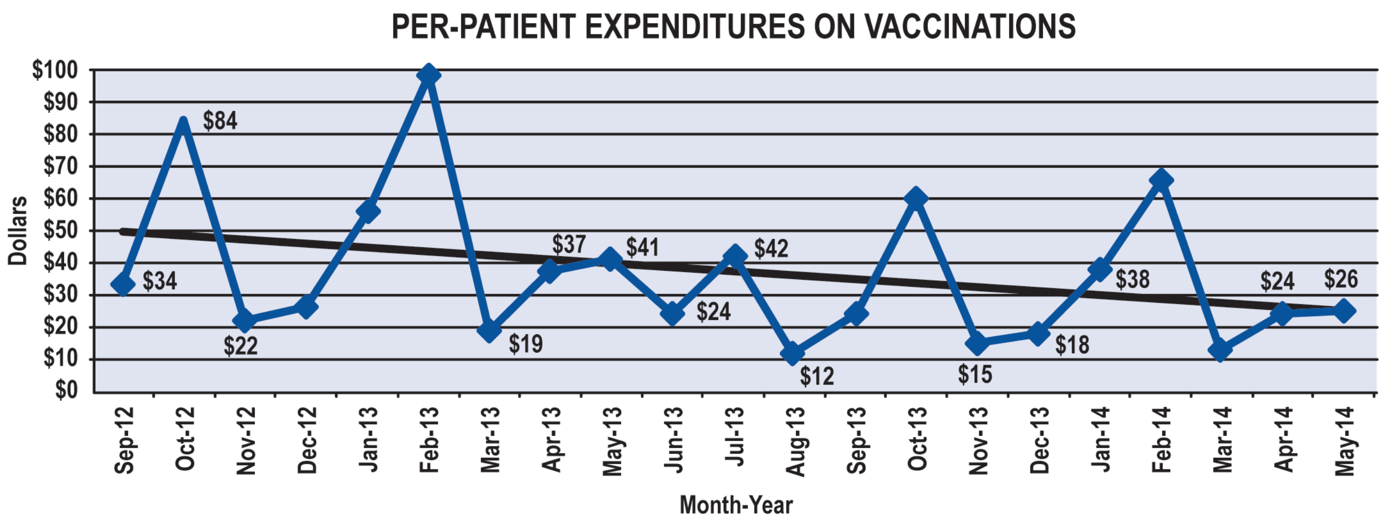
Figure 11. Attribute Cross-Tabulation Example

Level of Satisfaction	Service Waiting Time (Minutes)			Totals
	10 or Less	11 to 20	21 to 30	
Unsatisfied	27	37	56	**120**
Neutral	35	39	41	**115**
Satisfied	43	33	30	**106**
Totals	**105**	**109**	**127**	**341**

It appears that client satisfaction with services received falls as wait time increases.

Alternatively, if the data are true quantities, not counts in Yes-No categories, a line graph, a close relative of the histogram, is useful. Assume the analysis of vaccine expenditures shown in Figure 12.

Figure 12. Variable Analysis Graph

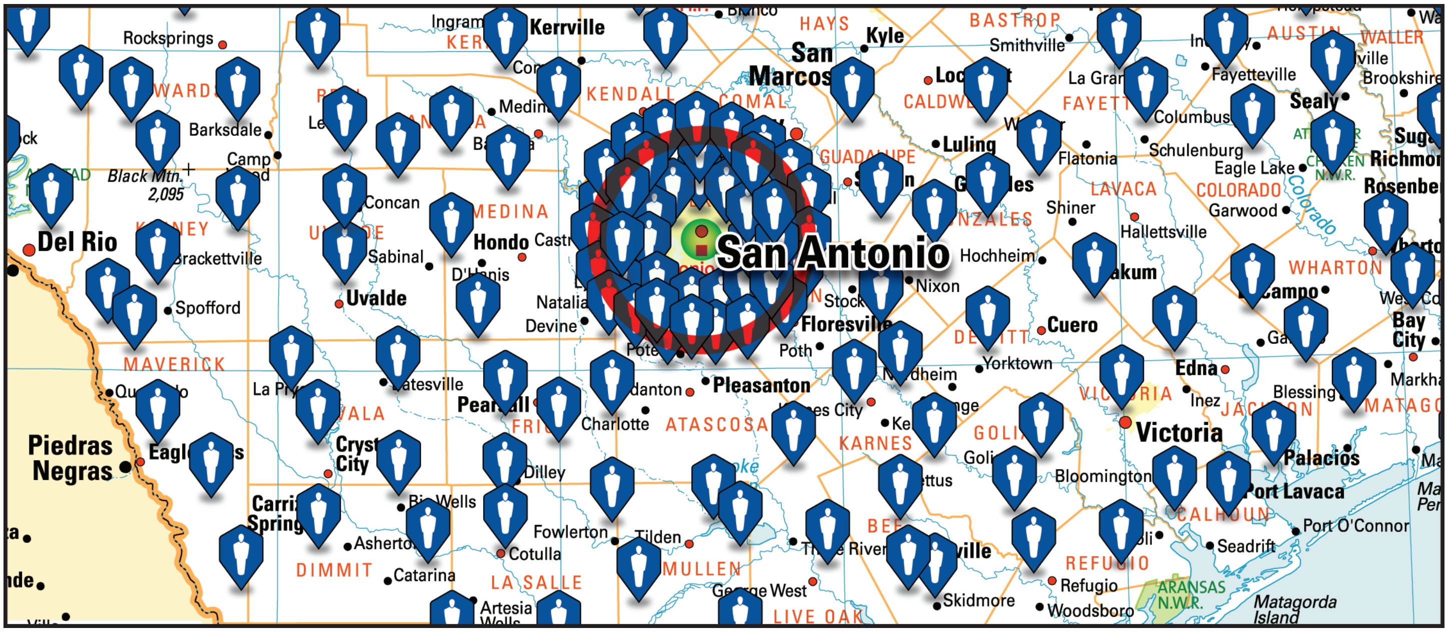


Expenditures seem to peak in October and February (seasonality) and trend (drift) downwards. Such data could influence staff allocation, vaccine purchase timing, or more aggressive marketing of vaccination services. Similar analyses can reveal trends in process stability or variation and possible process improvement opportunities.

17. Explore Distance

Maps are powerful summaries that can lend prominence to potential anomalies. For example, suppose that the map[23] in Figure 13 depicts distances traveled by Medicaid beneficiaries to fill prescriptions for a widely available controlled substance. The red circle defines a distance three standard deviations above the average distance from the pharmacy represented by the green dot.

Figure 13. Distance Analysis Map



The map prompts such questions as “why travel to that pharmacy if the drug is available at closer locations” and “are beneficiaries filling the prescription at multiple drug stores?” Global Positioning System (GPS) coordinates can be used for similar visual and statistical analyses of consumption and supply and other relationships between variables of interest. Information on GPS is available from http://www.gps.gov/ and free mapping software is available on the Internet.

The procedures discussed here generally apply to any health care transaction data set, as a function of the question(s) to be answered and the type of data that apply. While more complex questions require more sophisticated methods and tools, familiarity with and regular execution of these procedures should help develop a solid foundation for basic data mining and analysis.

To see the electronic version of this job aid and the other products included in the “Basic Data Mining and Analysis for Program Integrity: A Primer for Physicians and Other Health Care Professionals” Toolkit, visit the Medicaid Program Integrity Education page at https://www.cms.gov/Medicare-Medicaid-Coordination/Fraud-Prevention/Medicaid-Integrity-Education/edmic-landing.html

Follow us on Twitter #MedicaidIntegrity

References

1 United States General Accounting Office. Program Evaluation and Methodology Division. (1991, June). Using Structured Interviewing Techniques. GAO/PEMD-10.1.5 (p. 6). Retrieved April 17, 2015, from http://www.gao.gov/assets/80/76090.pdf

2 Centers for Medicare & Medicaid Services. Information Systems Security Awareness Training. Information Systems Security and Privacy Awareness Training Fiscal Year 2014 (slides 19–51). Retrieved April 17, 2015, from https://www.cms.gov/Research-Statistics-Data-and-Systems/CMS-Information-Technology/InformationSecurity/Downloads/Information_Systems_Security_and_Privacy_Awareness_Training.pdf

3 Centers for Medicare & Medicaid Services. Information Systems Security Awareness Training. Information Systems Security and Privacy Awareness Training Fiscal Year 2014 (slides 64–66). Retrieved April 17, 2015, from https://www.cms.gov/Research-Statistics-Data-and-Systems/CMS-Information-Technology/InformationSecurity/Downloads/Information_Systems_Security_and_Privacy_Awareness_Training.pdf

4 U.S. Department of Commerce. National Institute of Standards and Technology. (2013, October 30). Engineering Statistics Handbook (section 2.4.5.1). Retrieved April 25, 2015, from http://www.itl.nist.gov/div898/handbook/mpc/section4/mpc451.htm

5 United States Government Accountability Office. By the Comptroller General of the United States Government (2011, December). Auditing Standards 2011 Revision. GAO-12-331G (sections 6.56–6.73). Retrieved April 22, 2015, from http://gao.gov/assets/590/587281.pdf

6 United States General Accounting Office. Program Evaluation and Methodology Division. (1992, June). Quantitative Data Analysis: An Introduction. GAO/PEMD-10.1.11 (pp. 16–18). Retrieved May 3, 2015, from http://www.gao.gov/assets/80/76118.pdf

7 United States General Accounting Office. Program Evaluation and Methodology Division. (1992, June). Quantitative Data Analysis: An Introduction. GAO/PEMD-10.1.11 (p. 4). Retrieved May 3, 2015, from http://www.gao.gov/assets/80/76118.pdf

8 United States General Accounting Office. Program Evaluation and Methodology Division. (1992, June). Quantitative Data Analysis: An Introduction. GAO/PEMD-10.1.11 (pp. 38–45). Retrieved May 3, 2015, from http://www.gao.gov/assets/80/76118.pdf

9 U.S. Department of Commerce. National Institutes of Standards and Technology. (2013, October 30). Engineering Statistics Handbook (section 1.3.5.11). Retrieved April 21, 2015, from http://www.itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

10 U.S. Department of Commerce. National Institutes of Standards and Technology. (2013, October 30). Engineering Statistics Handbook (section 6.5.1). Retrieved April 21, 2015, from http://www.itl.nist.gov/div898/handbook/pmc/section5/pmc51.htm

11 United States General Accounting Office. Program Evaluation and Methodology Division. (1992, June). Quantitative Data Analysis: An Introduction. GAO/PEMD-10.1.11 (p. 22). Retrieved May 3, 2015, from http://www.gao.gov/assets/80/76118.pdf

12 Encyclopedia of Mathematics. Three-Sigma Rule. Retrieved April 27, 2015, from https://www.encyclopediaofmath.org/index.php/Three-sigma_rule

13 Centers for Medicare & Medicaid Services. Center for Medicaid and CHIP Services. CMCS Informational Bulletin. (2013, July 24). Targeting Medicaid Super-Utilizers to Decrease Costs and Improve Quality (p. 25). Retrieved May 1, 2015, from http://medicaid.gov/federal-policy-guidance/downloads/cib-07-24-2013.pdf

14 U.S. Department of Commerce. National Institutes of Standards and Technology. (2013, October 30). Engineering Statistics Handbook (section 7.1.6). Retrieved May 3, 2015, from http://www.itl.nist.gov/div898/handbook/prc/section1/prc16.htm

15 Centers for Medicare & Medicaid Services. (2013, January 11). Prescription Drug Benefit Manual. Chapter 9–Compliance Program Guidelines (section 50.6.9). Retrieved April 27, 2015, from https://www.cms.gov/Medicare/Prescription-Drug-Coverage/PrescriptionDrugCovContra/Downloads/Chapter9.pdf

16 U.S. Department of Commerce. National Institutes of Standards and Technology. (2013, October 30). Engineering Statistics Handbook (section 1.3.3.21). Retrieved May 3, 2015, from http://www.itl.nist.gov/div898/handbook/eda/section3/normprpl.htm

17 United States General Accounting Office. Program Evaluation and Methodology Division. (1992, June). Quantitative Data Analysis: An Introduction. GAO/PEMD-10.1.11 (p. 22). Retrieved May 5, 2015, from http://www.gao.gov/assets/80/76118.pdf

18 U.S. Department of Commerce. National Institutes of Standards and Technology. (2013, October 30). Engineering Statistics Handbook (section 1.3.3.14). Retrieved April 20, 2015, from http://www.itl.nist.gov/div898/handbook/eda/section3/histogra.htm

19 University of Victoria. Anscombe Data. Retrieved April 26, 2015, from http://web.uvic.ca/~dgiles/blog/anscombe.xls

20 Statistics How To. (2014, January 24). Misleading Graphs: Real Life Examples. Retrieved May 3, 2015, from http://www.statisticshowto.com/misleading-graphs/

21 United States General Accounting Office. Program Evaluation and Methodology Division. (1991, June). Using Structured Interviewing Techniques. GAO/PEMD-10.1.5 (p. 56). Retrieved May 4, 2015, from http://www.gao.gov/assets/80/76090.pdf

22 United States General Accounting Office. Program Evaluation and Methodology Division. (1992, June). Quantitative Data Analysis: An Introduction. GAO/PEMD-10.1.11 (p. 49). Retrieved April 29, 2015, from http://www.gao.gov/assets/80/76118.pdf

23 MapPoint Forums. Map Forums. MP2K Magazine Articles. Using MapPoint for Cluster Analysis: Part 1, Applications. Retrieved April 29, 2015 from http://www.mapforums.com/images/articles/MapPoint_Cluster1.png

Disclaimer

This job aid was current at the time it was published or uploaded onto the web. Medicaid and Medicare policies change frequently so links to the source documents have been provided within the document for your reference.

This job aid was prepared as a service to the public and is not intended to grant rights or impose obligations. This job aid may contain references or links to statutes, regulations, or other policy materials. The information provided is only intended to be a general summary. Use of this material is voluntary. Inclusion of a link does not constitute CMS endorsement of the material. We encourage readers to review the specific statutes, regulations, and other interpretive materials for a full and accurate statement of their contents.

November 2015