Basic Data Mining and Analysis for Program Integrity: A Primer for Physicians and Other Health Care Professionals

Presentation

Learning Objectives

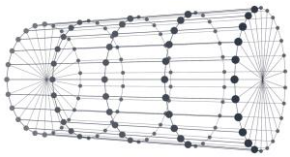
At the conclusion of this presentation, the learner will be able to:

- Recognize what data mining and analysis involve
- Recall the benefits and challenges of data mining and analysis
- Recall data mining and analysis models and analysis tools
- Identify sources of data that may be used in Medicaid data mining and analysis

What Is Data Mining?

Data mining is:

- Using a database to uncover data patterns and relationships and infer rules to predict future results
- Transforming data into actionable information



Data Mining for Program Integrity

Program integrity activities ensure that:

- Providers meet Federal and State participation requirements
- Services provided are medically necessary and appropriate
- Payments amounts are correct

Data Mining Processes

Three commonly used data mining processes are:

1. Knowledge Discovery in Databases (KDD)
2. Sample, Explore, Modify, Model, Assess (SEMMA)
3. CRoss Industry Standard Process for Data Mining (CRISP-DM)

Knowledge Check

Data mining processes:

A. Usually involve sequential, generally identifiable steps

B. Have no industry standard or universally defined approach

C. Develop models that uncover patterns or predict outcomes

D. All of the above

Benefits

Benefits of data analysis:

- Increased understanding of the business
- Improved decisions
- Identifying and analyzing risks and controls
- Building better systems
- Demonstrating commitment to program integrity and doing things right
- Creating the "sentinel" effect

Benefits—Continued

Makes us better stewards of a large national investment

2014 Medicaid and Medicare Summary Data				
Program	Clients Served	Amount Spent	Improper Payments	Estimated Percent Spent
Medicaid	65 million	$508 billion	$34 billion	6.7%
Medicare	54 million	$603 billion	$60 billion	10%

Benefits—Continued

Data mining and analysis can also help you:

- Prepare for, and respond to, an audit or review
- Better understand how the discipline of statistics works
- Assess progress toward predictive analytics
- Classify audit questions:
  - Descriptive (univariate)
  - Normative (bivariate)
  - Cause-and-effect (multivariate)

Challenges

Data mining and analysis challenges:

- Garbage in, garbage out
- Good data starts with security
- Understand, manage, and reduce the risk to information under the control of the organization
- Golden Rule principle

Challenges—Continued

Good data requires:

- Security, access, and privacy controls
- Policies, procedures, and training
- Firewalls, backup, and retention
- Optimum protection of personally identifiable information (PII) that can distinguish or trace an individual's identity

Knowledge Check

All data are good data.

A. True
B. False

Data Quality

Before mining or analyzing data:

- Understand and vet data and its dictionary
- Use a unique row counter
- Save original data, and use a copy
- Document what you want to know
- Delete irrelevant fields or records

In general, when addressing data quality, focus on:

- Finding and disposing of errors in the data
- Finding and disposing of bias in the data

Data Errors

- Does the error relate to the question(s) asked?
  - If so, how?
- Does it matter?
- Need other data?
- Change question(s) asked?
- Raise other or future issues?
- Look for things like blanks, error codes, inappropriate zeroes, unreasonable values, etc.

Addressing Bias

Bias is the:

- Difference between what is and what ought to be
- Foundation of program integrity and fraud

The system and analytical approach must:

- Detect and show small changes
- Respond to change correctly
- Show slow change over time
- Specify differences between high-side and low-side values
- Be consistent with like systems
- Reveal variances in process, policy, practice, or people

Basic Data Mining Models

There are four groups of useful data mining models:

1. Rules-based
2. Anomaly detection
3. Predictive
4. Network analysis

Rules-Based Models

Rules-based models include:

- Algorithms
- National Correct Coding Initiative (NCCI) edits
- Medically Unlikely Edits (MUEs)

Anomaly Detection

Anomaly detection tools for data mining include:

- Descriptive statistics
- Outlier analysis
- Data visualization
- Correlation
- Peer comparison
- Pattern analysis and trending
- Geographic and map analysis

Anomaly Detection—Descriptive Statistics

Most program integrity work focuses on:

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.

Anomaly Detection—Descriptive Statistics—Continued

Normal Distribution:

- Bell curve as criteria
- Most important and frequent distribution

Abnormal Distribution:

- Mean, median, mode differ
- Data spread out
- Data concentrate
- Data asymmetrical

Anomaly Detection—Outliers

An outlier is "an observation that appears to deviate markedly from other observations." That is, an item significantly different from, quantitatively or qualitatively, the norm or comparative data set.

Anomaly Detection—Outliers—Continued

Deal with outliers by:

- Investigating them carefully to obtain information on the process or the data
- Determining their source(s) and reason(s) for their appearance
- Finding out if they are simply "bad data"

Possibilities include:

- Insensitive metrics to outliers, such as the median (rather than the mean)
- Eliminating them from the analysis
- Treating them as a separate group

Anomaly Detection—Data Visualization

Graphs are the basic form of data visualization.

- Graphs can show:
  - Central tendency
  - Spread
  - Asymmetry
  - Possible outliers
  - Multiple modes

Anomaly Detection—Correlation

- Measures linear relationships
- Tells us how much and in which direction one value changes when another value changes
- Varies between -1 and +1
- Zero means no relationship

Consider strength and direction of the correlation value in each cell and consider:

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

Correlation is not causation.

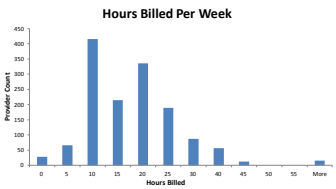
Anomaly Detection—Peer Comparison

- To compare attributes, look at percentages
- Focus on differences between categories
- The example suggests Hilltop bills late; Valley View does better

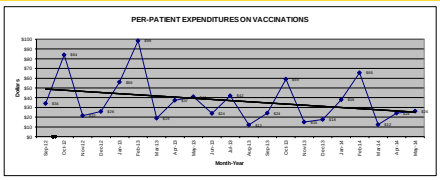
Billing Clinic	Payment Aging			Totals
	0-30 Days	30-60 Days	Over 60 Days	
Hilltop Clinic	23%	31%	46%	100%
Woodlands	30%	34%	36%	100%
Valley View	41%	31%	28%	100%

Anomaly Detection—Peer Comparison—Continued

- To compare quantitative variables, look at numbers
- Focus on differences in numerical values
- The example reveals a possible outlier at the far right



Anomaly Detection—Pattern Analysis and Trending



- To analyze attributes or ranks, consider "cross-tabs"
- To analyze variables, consider histograms or line graphs
- The example suggests downward drift
- Can guide resource allocation and use

Anomaly Detection—Geographic and Map Analysis

Maps make anomalies prominent.

For example, the map shows the distances traveled by Medicaid beneficiaries to fill prescriptions.

Predictive Models

Predictive models include:

1. Decision tree analysis
2. Predictive modeling
3. Cluster analysis

Network Analysis Models

Network analysis models include:

1. Social network analysis
2. Link analysis
3. Association analysis

Knowledge Check

Predictive models typically describe the likelihood of behavior patterns among people or objects and network models typically uncover relationships among people or objects.

A. True

B. False

Data Sources

Data sources used for, or arising from, data mining for program integrity fall into three groups:

1. Provider practice data
2. State data
3. Federal data

Data Source—Provider Practice Data

Provider practice data sources:

- Medical billing and compliance software
- Electronic health records (EHR) system

Data Sources—State Level

Primary source:

- The primary data source for most State Medicaid programs is the State Medicaid Management Information System (MMIS)

Data Sources—State Level —Continued

Other sources

- State Program Integrity Unit (PIU)
- Medicaid Fraud Control Unit (MFCU)
- Medicaid Recovery Audit Contractor (RAC)
- State auditor, comptroller, or treasurer
- Claims administrator
- Fiscal agent
- Managed care organizations
- Professional associations

Data Sources—Federal Level

Primary sources

- U.S. Department of Health and Human Services, Office of Inspector General (HHS-OIG)
- Centers for Medicare & Medicaid Services (CMS)

Data Sources—Federal Level —Continued

Other sources

1. CMS Healthcare Fraud Prevention Partnership (HFPP)
2. Medicare Administrative Contractor (MAC)
3. Program Safeguard Contractors (PSCs) and Zone Program Integrity Contractors (ZPICs)
4. Comparative Billing Reports (CBRs)

Federal Data Sources for Oversight

Sources include:

- National Plan & Provider Enumeration System (NPPES)
- Provider Enrollment, Chain and Ownership System (PECOS)
- List of Excluded Individuals/Entities (LEIE)
- Excluded Parties List System (EPLS) on the System for Awards Management (SAM)
- Other State and Federal debarments
- CMS Data Navigator

Knowledge Check

The three broad sources of data that may be used in Medicaid data mining and analysis are:

A. Provider data, World Health Organizational data, and American Medical Association data

B. State data, Federal data, and World Health Organizational data

C. Provider data, State data, and Federal data

D. American Medical Association data, State data, and Federal data

Conclusion

In this presentation we have:

- Defined data mining and analysis
- Explained the benefits and challenges of data mining and analysis
- Described various data mining models and data analysis methods
- Identified sources of data that may be used in Medicaid data mining and analysis

Questions

Please direct questions or requests to: MedicaidProviderEducation@cms.hhs.gov.

To see the electronic version of this presentation 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 on the CMS website.

Follow us on Twitter #MedicaidIntegrity

Disclaimer

This presentation 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 presentation was prepared as a service to the public and is not intended to grant rights or impose obligations. This presentation 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