American Health Information Management Association

HEALTH DATA ANALYSIS TOOLKIT

AHIMA
American Health Information Management Association®

HEALTH DATA ANALYSIS TOOLKIT

ISBN: 978-1-58426-442-2

AHIMA Product No.: ONB194014

AHIMA Staff:

Jessica Block, MA, *Project Editor*
Pamela Woolf, *Director of Publications, AHIMA Press*
Anne Zender, *Senior Director, Periodicals*

For more information about AHIMA Press publications, including updates, visit ahima.org/education/press.

American Health Information Management Association
233 N. Michigan Ave., 21st Fl.
Chicago, Illinois 60601

TABLE OF CONTENTS

Foreword 4
Authors and Acknowledgments 4
Introduction 5
Data Dictionary 7
Study Design and Report Request Form 9
Data Capture Methods and Best Practices 10
Validating Data Outcomes 11
Presenting the Data 13
Formulas and Statistics 19
Sample Job Skills and Responsibilities 28
Glossary of Terms 32
Annotated Bibliography 41
Case Study 48
Appendix A: Data Dictionary Sample 49
Appendix B: Study Report and Request Form 50
Appendix C: Sample Data Definitions 52
Appendix D: Meaningful Use 62
Appendix E: Clinical Quality Measures 64

FOREWORD

This toolkit provides a variety of resource tools for healthcare professionals performing data analysis tasks, whether they are reviewing and trending healthcare data for a healthcare entity, reporting on quality measures in a physician office, managing the enterprise master patient index, or working with other operational and financial data. The toolkit begins by discussing the current healthcare initiatives and programs that are increasing the demand for data analytics. From there, the resources focus on the acquisition of the data and the purpose of the data dictionary and include resources from the point when information is requested via a sample report request form to how data should be validated. Examples of good and bad data displays are discussed. The second half includes a listing of common formulas and statistics, a glossary of terms, and an annotated bibliography used in data analysis. Finally, the toolkit concludes with a sample case study illustrating how data are collected, analyzed, and transformed into information for reporting.

AUTHORS

Julie Dooling, RHIA, CHDA

Pawan Goyal, MD, MHA, MS, PMP, FHIMSS, CPHIMS, CBA

Linda Hyde, RHIA

Lesley Kadlec, MA, RHIA

Susan White, PhD, RHIA, CHDA

ACKNOWLEDGEMENTS

Michelle Custodio, RHIA, CCDS, CDIP

Jane DeSpiegelaere, MBA, RHIA, CCS, FAHIMA

Katherine Downing, MA, RHIA, CHPS, PMP

Kim Turtle Dudgeon, RHIT, CHTS, IS/TS, CMT

Angela Dinh Rose, MHA, RHIA, CHPS, FAHIMA

Carol F. Smith, MBA, RHIA, FAHIMA

Diana Warner, MS, RHIA, CHPS, FAHIMA

Lou Ann Wiedemann, MS, RHIA, FAHIMA, CDIP, CHDA, CPEHR

ORIGINAL AUTHORS

June Bronnert, RHIA, CCS, CCS-P

Jill S. Clark, MBA, RHIA

Linda Hyde, RHIA

C. Jeanne Solberg, MA, RHIA

Susan White, PhD, RHIA, CHDA

Mark Wolin, MHSA

ORIGINAL ACKNOWLEDGEMENTS

Jane DeSpiegelaere-Wegner, MBA, RHIA, CCS, FAHIMA

Angela K. Dinh, MHA, RHIA, CHPS

Kathy Giannangelo, MA, RHIA, CCS, FAHIMA

Syreeta Kinnard, RHIA

Wendy Scharber, RHIT, CTR

Michelle Shimmel, RHIT

Mary H. Stanfill, MBI, RHIA, CCS, CCS-P, FAHIMA

Diana Warner, MS, RHIA, CHPS

Lou Ann Wiedemann, MS, RHIA, FAHIMA, CPEHR

Kelli Wondra

INTRODUCTION

As electronic health record (EHR) use advances, the available data elements continue to expand. Healthcare has become a data-rich field. Various government initiatives, such as the transition to the International Classification of Diseases 10th Revision Clinical Modification/Procedure Coding System (ICD-10-CM/PCS) and legislation such as the American Recovery and Reinvestment Act and the Patient Protection and Affordable Care Act, further enhance the necessity for organizations to analyze their data to make informed business decisions. This toolkit provides a resource that facilitates and supports data analytics performed by health information management (HIM) and other professionals at a foundational level.

VALUE-BASED PURCHASING

Healthcare payers are moving from payment for volume to pay for value and performance. Data analysis can help healthcare organizations maximize their value as providers to patients and payers. The first phase of the CMS value-based purchasing (VBP) program was implemented in October 2012. Hospitals paid via the inpatient prospective payment system (IPPS) found two significant changes to their payment during FY 2013. First, implementation of the VBP program reduced payment by 1 percent, but provided hospitals with the opportunity to earn back that 1 percent reduction and more by performing well according to the CMS Total Performance Score (TPS).[1] Second, some hospitals faced reductions in payment due to the Readmissions Reduction Program.

Data mining to determine which VBP metric is causing a reduction in a hospital's payment requires not only knowledge of healthcare data analytics, but also knowledge regarding reimbursement and the business side of healthcare. Quality measurement requires attention to data quality and validity. Health data analysts are needed to design sampling plans for abstracted measures and specify data extract parameters for administrative data-driven measures. All of these roles create new opportunities for HIM professionals that can bring both context and content to the table in healthcare data analytics.

MEANINGFUL USE AND CLINICAL QUALITY MEASURES

In order for an eligible professional or hospital to attest to having met the objectives of the Medicare and Medicaid EHR incentive programs, specific data must be collected on each provider and hospital for a specified time period, and a percentage is calculated based upon a specific subset of patients. There are inclusion and exclusion criteria for each objective. These incentive programs provide financial incentives for the "meaningful use" of certified EHR technology to improve patient care. The ability to demonstrate "meaningful use" is driven by electronic data capture, management, analysis, and reporting. Health data analysts play a key role in obtaining the necessary data and analyzing it to ensure it is correct.

For more details on Meaningful Use, please see Appendix D.

In addition to meeting the core and menu objectives, eligible professionals, eligible hospitals, and critical access hospitals are also required to report clinical quality measures. For more details, please see Appendix E.

1. Centers for Medicare and Medicaid. "Frequently Asked Questions, Hospital Value-Based Purchasing Program." March 9, 2012. http://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/hospital-value-based-purchasing/Downloads/FY-2013-Program-Frequently-Asked-Questions-about-Hospital-VBP-3-9-12.pdf.

HOW DATA ANALYTICS ALIGNS WITH AHIMA STRATEGIC INITIATIVES

Within the healthcare environment, data is limitless. Whether it is clinical, administrative, financial, patient-generated or other data, the need to manage the data efficiently is more important than ever.

As healthcare organizations (HCOs) continue to implement EHRs and new care delivery models such as accountable care organizations (ACOs) flourish and grow, demands for accountability, reliability, and security must be met.

Now is the time for healthcare professionals to start creating and participating in *information governance* programs. In healthcare, the introduction of information governance represents something of a change, though for many other industries information governance programs have been quite successful. Information governance is a strategic initiative for AHIMA.

The goal of an information governance program is to ensure that all information resources support the business goals of an organization. Health data is at the core of information governance. Having the knowledge to acquire, manage, analyze, interpret, and transform this data into accurate, consistent, and timely information, while balancing the strategic vision with day-to-day details, is a core responsibility of a health data analyst and is a key factor in the success of information governance programs.[2]

WHAT IS DATA?

Data is plural of datum, which is the dates, numbers, ages, symbols, letters, and words that represent basic facts and observations about people, processes, measurements, and conditions.[3]

Data could be qualitative or quantitative. (See the glossary of terms.) It could be collected and represented as an individual or set of numbers, alphabets, symbols, pictures, voice, or video.

Data in its raw form is of limited value. It needs to be processed in a meaningful way to create information that would be relevant to a situation. When the information is used for decision making, it leads to knowledge creation for common understanding. Repeatable use of knowledge leads to development of best practices which, applied over time to achieve organizational goals, leads to behavior change.

2. Thomas Gordon, Lynne. "Information Governance for the Health Care Industry: Now is the Time." *iHealthbeat*, Feb. 3, 2014. http://www.ihealthbeat.org/perspectives/2014/information-governance-for-the-health-care-industry-now-is-the-time
3. AHIMA. *Pocket Glossary of Health Information Management and Technology,* 4th ed. Chicago: AHIMA Press, 2014.

DATA DICTIONARY

A data dictionary is a descriptive list of names (also called "representations" or "displays"), definitions, and attributes of data elements to be collected in an information system or database. The purpose of the data dictionary is to standardize definitions and ensure consistency of use. It is a tool to aid in the standardization of data definitions. It is not and should not be confused with other terms such as a "pick list" or "drop-down menu" within an EHR. See Appendix A for a sample data dictionary.

A key focus of a data dictionary is to support and adopt more consistent use of data elements and terminology to improve the use of data in reporting. A data dictionary promotes clearer understanding, helps users find information, promotes more efficient use and reuse of information, and promotes better data management. The data dictionary is a critical component of data governance.

According to the International Organization for Standardization:

> The increased use of data processing and electronic data interchange heavily relies on accurate, reliable, controllable, and verifiable data recorded in databases. One of the prerequisites for a correct and proper use and interpretation of data is that both users and owners of data have a common understanding of the meaning and descriptive characteristics (e.g., representation) of that data. To guarantee this shared view, a number of basic attributes has to be defined.[4]

A dictionary describes the definitions or the expected meaning and acceptable representation of data for use within a defined context of data elements within a data set. In addition to the name and definition of the data element, the metadata may include other attributes or characteristics such as length of data element, data type (character or numeric), data frequency (mandatory or not), allowable value and constraints, originating source system, data owner, data entry date, and data termination date.

Data are often stored in many different databases and may be of variable quality. Inconsistent naming conventions, inconsistent definitions, varying field length for the same data element, and/or varied element values all can lead to significant problems, including poor data quality and misuse of data in reporting, among others. The following are a few of many examples of inconsistent data throughout an organization:

1. Inconsistent naming conventions
   - The date of the patient's admission is referred to as the *date of admission* in the patient management (PM) system, *admit date* in the fetal monitoring system, and *admission date* in the cardiology database.
   - The unique patient identifier is referred to as a *medical record number* in the PM system, *patient record identifier* in the operating room system, and *"A" number* (a moniker leftover from a legacy system from 25 years ago) and *enterprise master patient identifier* in the catheterization laboratory system.
2. Inconsistent definitions
   - Admission, discharge, transfer (ADT) system: *date of admission* is the date on which an inpatient or day surgery case admission occurs; in the trauma registry system, *date of admission* is the date on which the trauma patient enters the operating room.
   - The *pediatric age* is defined as age less than or equal to 13 in the PM system, whereas the pediatric disease registry defines a *pediatric age* as less than the age of 18.
   - In the bed board system, a *nursing unit* may be defined as 5W or 5 West. Within the scheduling system, *unique locations* are defined as *short procedure unit* or *SPU,* such as *X-ray* or *radiology,* for example.

4. International Organization for Standardization, 2004. "Information Technology, Parts 1–6." (2nd ed). www.iso.org/iso/home.html.

3. Varying field length for same data element
   - The field length for a patient's last name is 26 in the PM system, whereas the field length for a patient's last name is 15 in the cancer registry system.
   - The medical record number in the PM system is 16 characters long, whereas the cancer registry system maintains a length of 13 characters for the medical record number.
4. Varied element values
   - The patient's sex is captured as *M, F,* or *U* in the ADT system, whereas the patient's sex is captured as *Male, Female,* or *Other* in the peripheral vascular laboratory database.

Data dictionaries facilitate the work of the health data analyst by developing a common understanding of an organization's data quality. Therefore, although ability to edit the organization's data dictionary should be limited to system administrators, it is important that the dictionary be viewable for all those who use data to manage their work, including but not limited to health data analysts. The dictionary is often organized as a table and should be in a format that others can access. To review the varying formats, please see Appendix A or use the following links:

- Research Data Assistance Center
- Medicaid and CHIP Statistical Information System
- Google image search on "data dictionary example"

STUDY DESIGN AND REPORT REQUEST FORM

Reports can be requested for a variety of reasons; therefore, it is imperative to define the objectives and parameters of the request—whether it is for a specific report, research study, or other type of analysis. The scope of the request will determine the amount of detail needed to ensure that the request can be completed successfully and the analysis will achieve the desired results. Before designing and running a requested report, analysts should document and review the requirements with the report's requester or end user. Doing so is critical to providing accurate information and avoiding rework. Below are a list of guidelines to consider when developing a report request form or designing a study.

1. Defining the study or report objectives
    a. If multiyear data are requested, will trends be analyzed? If so, then the consistency and completeness of the data for the entire period of the study must be considered.
    b. If multisite data are requested, thresholds need to be set to determine the eligibility of each site for the request. If hospital characteristics are important to the study (e.g., teaching, size, geographic location), then distribution of sites and data according to these characteristics needs to be considered when determining the final study population.
    c. What requirements are needed to meet privacy and security requirements for requests involving patient-level data? Will the data set needed for analysis qualify as a limited data set? Will institutional review board approval be needed? If the data are not already de-identified, what method will be used to mask protected health information data elements?
2. Identifying the correct study or report population
    a. How will the population be identified for the request?
    b. What are the sources of the data used to fulfill the request? If data are needed from multiple sources, how will the data be linked or merged for the final analysis data set?
    c. What types of coded data will be needed (e.g., ICD-10-CM, CPT, LOINC, SNOMED CT, RxNorm, UB codes)?
    d. Are there any known limitations in the data set that would affect the interpretation of the results of the report or analysis?
3. Determining the calculations and statistical tests needed for the request
    a. Will additional derived data fields be needed? What raw data will be needed to create the new data fields?
    b. What types of descriptive and inferential statistics will be used? Define the software program used to create the statistics and which tests will be used.
4. Designing the presentation
    a. How will the data be presented (charts, graphs, tables)?
    b. What format will be used for data presentation (electronic media, hard copy)?

For more guidance on designing a data request report form, see Appendix B.

DATA CAPTURE METHODS AND BEST PRACTICES

With the advent of new electronic data capture methods such as mobile health apps, patient self monitoring, patient portals, and health information exchange, there has been a dramatic increase in the ways that healthcare organizations acquire patient data for use in the EHR.

The patient's administrative and clinical data is captured at various sources inside and outside the healthcare organization. The process is typically initiated by registering the patient. At the point of registration, data such as the patient's demographic and insurance information is captured. As the patient moves throughout the visit, additional data is collected at each care site.

Capturing electronic discrete data elements can include the following: pre-defined or custom-built templates, electronic forms with or without drop-down menus, use of bar coding technology, direct entry into free text fields, front-end or back-end speech recognition with or without applied natural language processing, traditional dictation, and transcription. Unstructured data is also captured such as handwritten notes and scanned images.

External data is captured and brought into the health record as well. One trend that is on the rise is patient-generated data, generated when patients enter their data into the healthcare provider's system via templates, drop-down menus, or free text fields.

Once the data is collected, many tools may be used to enhance the capture process. Examples include the use of optical character recognition (OCR) software to read and parse data, such as in the remittance advice process or the use of natural language processing (NLP) or natural language understanding (NLU) in computer-assisted coding (CAC) or transcription processes.

Data workflows vary depending upon the care setting. The workflow in physician practice is more streamlined when compared to the hospital outpatient or inpatient processes. In all care settings, charges are captured through a coding classification system such as ICD-9-CM/PCS, CPT, or HCPCS. The data is then transferred to a claim form such as the UB-04 for inpatient or outpatient or the CMS-1500 for professional fees.

Best practices for EHR data capture should incorporate the following:[5]

- Consider what data needs to be captured and customize available tools to collect it.
- Evaluate the data and determine its placement in the record to determine what rules or procedures need to be put in place to upload the information most efficiently and without errors.
- Collect the data in a standardized format using templates or discreet fields to make retrieval for reporting easier.
- Routinely audit a sample of records that were collected using the data capture methods described above.
- Acquire primary and secondary data from existing internal or external data sources.

5. White, Susan. *A Practical Approach to Analyzing Healthcare Data*. Chicago: AHIMA Press, 2013.

VALIDATING DATA OUTCOMES

Validation of data can occur at many different points in the capturing, storing, managing, reporting, and analyzing process. The validation steps outlined below focus on validating the outcomes during the reporting and performing of analysis tasks.

A. Data Extraction and Aggregation

1. Identify the source of all the data elements for the request.
2. When a data dictionary is available for the data elements being used, become familiar with how the raw data were collected and identify any potential issues with the data. For example, if your analysis is based only on Medicare patients, are you identifying these patients by using a data field that is complete and up to date?
3. If data are being merged or linked across multiple sources, create a plan to ensure that the linkage is correct. If there is a need to access laboratory results from the laboratory information system and link to patients in the billing system, what method will be used to match the patient and encounter information? Create a frequency of matched patients to determine that the match rates make sense (e.g., should all patients in your analysis population have at least one laboratory test).
4. For classification systems used to aggregate data, verify that the codes selected are appropriate for the time frame of the study. If the analysis period crosses times when the classification system has changed, there may be codes that are valid for only part of the period. For example, ICD codes are updated every October 1. If your study spans a calendar year, you may need to include codes that appear for discharges only in the last quarter of the year that are consistent with the codes that were used in the first three quarters of the year.

B. Calculations and Statistics

1. When creating derived data fields, the analyst must determine a method to verify the results of the calculations. Derived fields must be reviewed before performing any further statistical tests.
2. Create a plan for verifying statistics used in the request.
   a. Do individual row percentages sum to 100?
   b. Do individual row counts sum to the grand totals?
   c. For complex data sets or statistical tests, determine a method to verify the results. Is there another staff member who can review program code or output, or are there other means to determine that tests were performed correctly?
   d. Identify references, such as reports from other departments for similar types of data or literature searches on the topic being analyzed that can be used to verify results. For example, if you are analyzing costs per member per month (PMPM) stratified according to service level, check the total PMPM against what is being reported by finance for the same or a similar period. If you are studying a particular disease or condition, reviewing published literature or external databases for outcomes such as mortality rates, length of stay (LOS), or readmissions can serve as a reference point in validating the results.

C. Presentation of Data

1. Ensure data are represented in accordance with request. Options for presentation may include but are not limited to:
   a. Pie chart
   b. Line graph
   c. Bar graph
2. Verify that all labels on charts and graphs are clear and readable to the user.
3. Compare the graph back to the source data used to create the graph, and make sure that the graph correctly represents the data.
4. Provide any additional information regarding inclusions or exclusions in a legend or other notation.
5. Verify all file names, figures, and tables against any documentation produced for the analysis.

Validating data outcomes when fulfilling a data request, including understanding the source of the data, verifying the calculations needed, and selecting the proper presentation methods, is critical to performing data analysis. Data display examples will be explored further in the following section.

PRESENTING THE DATA

This section will touch on data analysis tools and highlight examples of poor data display.

PIVOT TABLES

Pivot tables are an excellent Excel tool to summarize data according to categories. For example, they may be used to summarize charges according to department, counts of coded data elements, etc. Pivot tables also provide flexibility for the end user or analyst to organize and filter the data in various ways before finalizing the analysis. Below is a small example of a pivot table totaling diagnosis codes according to sex:

	Count of Dx		
Dx	Female	Male	Grand Total
4242		1	1
4254	1	1	2
4280	1		1
4290	1		1
5119		1	1
7455	1		1
78341	1		1
7861	1		1
Grand Total	7	3	10

The following resources offer instruction on creating Microsoft pivot tables:

Microsoft resources:

- http://office.microsoft.com/en-us/excel-help/pivottable-reports-101- HA001034632.aspx
- http://office.microsoft.com/en-us/excel-help/pivottable-i-get-started-with-pivottable- reports-in-excel-2007-RZ010205886.aspx

Non-Microsoft resources:

- http://www.brighthub.com/computing/windows-platform/articles/27415.aspx
- http://www.youtube.com/watch?v=7zHLnUCtfUk
- http://www.youtube.com/watch?v=i67XK3qjL_w

CREATING A FREQUENCY TABLE

Frequency tables are another useful Excel tool to summarize a set of data, specifically by recording how often each value (or set of values) occurs. As an added benefit, frequency tables can use percentages to further enhance the results. Below are instructions for how to create a frequency data table by using raw diagnostic or procedure data.

1. Place all diagnostic or procedure data into column A in an Excel spreadsheet.
2. Click on the cell to the right of the first number in the column.
3. On the Formula menu, choose More Functions, Statistical and then COUNTIF.
4. For the Range, highlight the entire group of cells that contain your data.
5. For the Criteria, highlight the first number in the column. This places the count of the number of times this code is present in the data into the cell where the pointer was located.
6. To copy this formula to count the remainder of the data, first enter a $ before the column letter and row number in both cell addresses. For example, if the address looks like this: A2:A4000, make the range look like this: $A$2:$A$4000 *(shortcut: highlight entire formula [A2:A4000] and then select F4)*.
7. Next, copy this formula into the cell just to the right of each number to be counted (in this case, from A3 though A4000). This formula counts how many times each code is present, but the count is listed as the result of a formula.
8. To translate this to a value, copy all of the formula results (highlight and Control C). Then, use Paste Special from the Home menu and click on Values, which copies the formula results into the same location but as actual numbers that can be sorted or calculated.
9. Highlight the entire spreadsheet by clicking on the upper left corner. On the Data tab, in the Data Tools group, click on Remove Duplicates. Uncheck everything but column A in the selection window and click OK.

This procedure results in a listing of each code present in the data and the corresponding count from the original data. Below is a small example of a frequency table created for diagnosis codes. You can sort these entries by either column A for number order or column B for frequency order.

Dx	Frequency
0092	1
135	1
243	4
311	3
319	12
0380	1
0382	1
0388	3
0389	10

OTHER RESOURCES

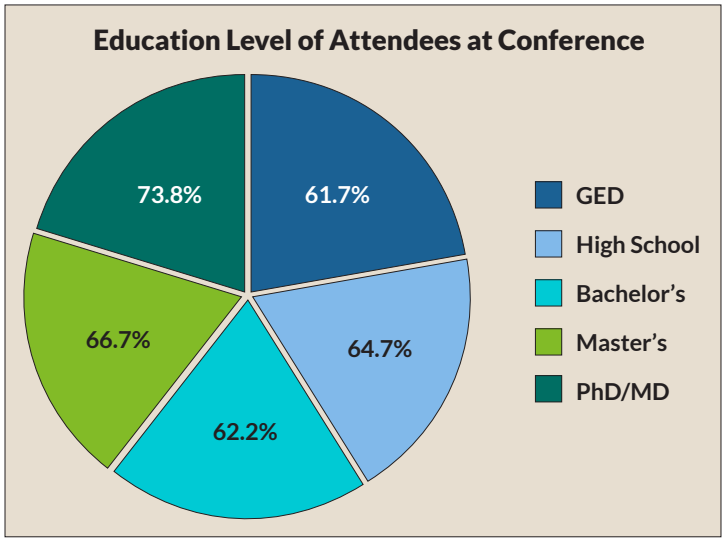
For more information on data analysis tools, such as statistical calculators, please refer to the "Calculations and Statistics" section of the Annotated Bibliography on page 41.

EXAMPLES OF POOR DATA PRESENTATION AND SUGGESTED ALTERNATIVES

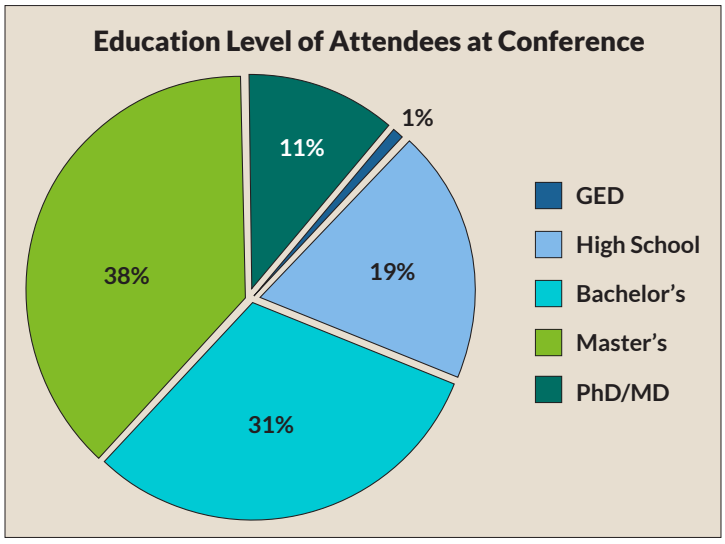
Numbers are factual data. However, when displayed incorrectly, they can be deceiving. Incomplete or misrepresented data may leave the reader with unanswered questions; therefore, healthcare data analysts must choose graphic representations that are appropriate to the information they are presenting. They should follow established guidelines for data visualization, such as using pie charts only for data that add up to a meaningful total. When displaying counts according to category, the categories being compared should be the same size; the size of the image displayed should be related to the corresponding data value; and data values must match other components of the graphic display. In the examples that follow, poor practice has been identified and guidance provided for improved data display.

EXAMPLE 1

Pie charts should be used only for data that add up to a meaningful total. In this first example, the sections do not.

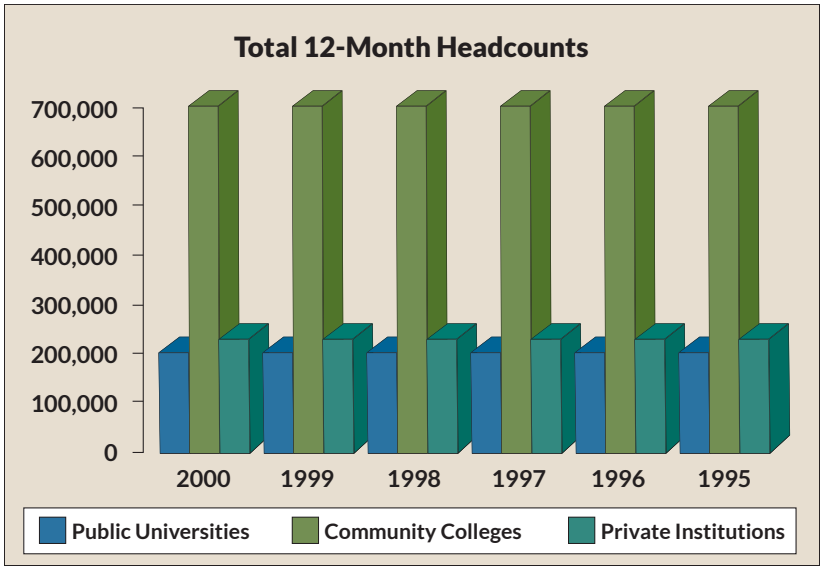


In comparison, the component parts of the pie chart below total 100 percent of the whole.

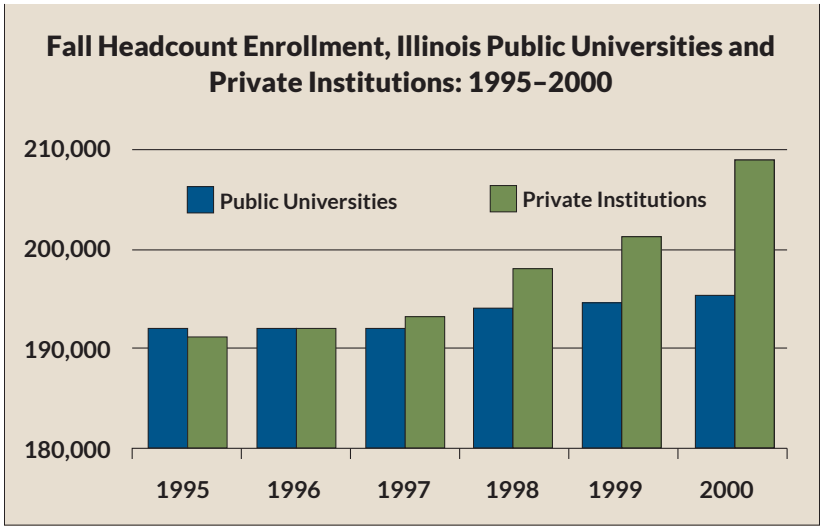


Example 2

In this example, time is displayed backward on the *x*-axis, and the three-dimensional effect is difficult to interpret: does the scale match the front or the back of the bars? Furthermore, the high values on the community college scale minimize the important variation in the other two series.



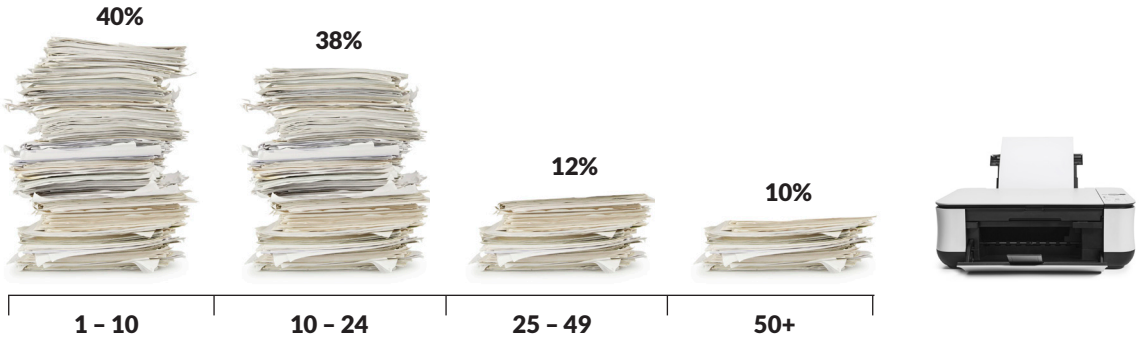
Below is a sample of how the same data could be displayed differently, omitting community colleges.



Example 3

In this example, the bin width is not consistent across the samples (1 to 10 versus 10 to 24).

Not Yet Paperless
The majority of office workers print up to 24 pages a day.

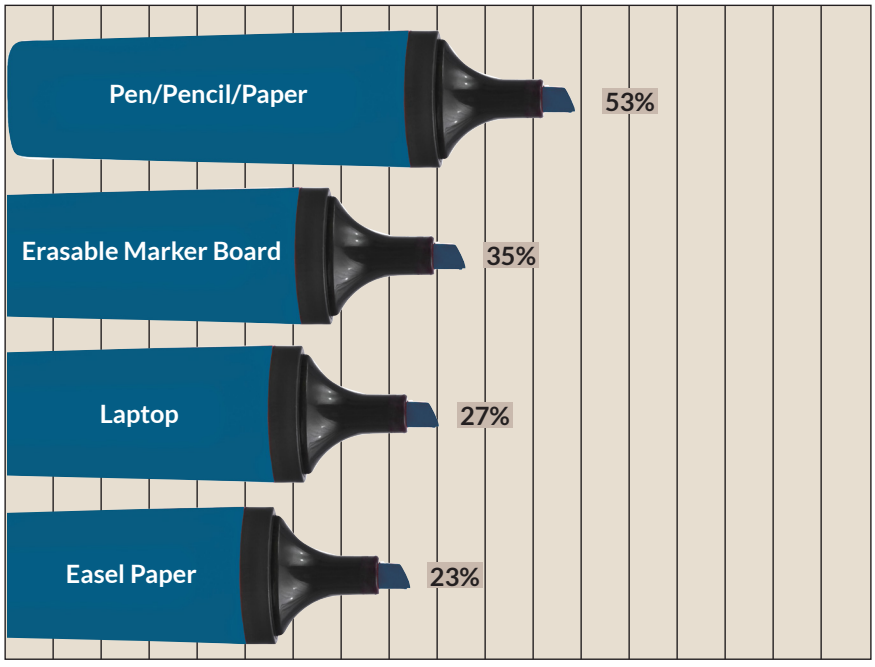


When displaying counts by category, the categories being compared should be the same size.

Example 4

Here, the proportionality is off—the marker representing "easel paper" is not half the length of the marking representing "pen/pencil/paper."

Pens Still Popular
More than half of workers take meeting notes with pen and paper.



Health data analysts must always ensure the size of the image displayed is related to the corresponding data value. For example, if the symbols are different widths, then the width should have an interpretation. In the example, even if the symbols had the same width, the length of the pencil representing "easel paper" should be half as long as the "pen/pencil/paper" symbol.

PREDICTIVE MODELING

Predictive modeling applies statistical techniques to determine the likelihood of certain events occurring together.[6] Statistical methods are applied to historical data to "learn" the patterns in the data. These patterns are used to create models of what is most likely to occur.

Predictive modeling is used by credit card issuers to determine if transactions are likely fraudulent. Customers who receive a phone call from their credit card company verifying that they authorized a transaction were the subjects of a predictive model.

For example, a customer's typical credit card transaction is $100. The credit card issuer notices that the customer submitted three $5,000 transactions in one day. Given the customer's history and the credit card issuer's historical data regarding fraudulent transactions, those transactions look suspicious.

The credit card company may then put a hold on the card and call to verify that the customer really did authorize the suspect transactions. The triggers that tell the credit card company when to suspect a fraud issue are created via predictive modeling techniques.

Predictive modeling techniques use multiple data sources. Data such as the provider's claim history, the patient's demographics and health status, the services included on the claim, and the attributes associated with previously identified fraudulent claims may all be used to develop a statistical model.

Statistical techniques used to create the model may include logistic regression, cluster analysis, or decision trees. All of these statistical techniques allow the user to combine multivariate historical data into a model that may be used to assess the probability or likelihood that current claims are fraudulent.

In logistical regression, the likelihood that a claim is fraudulent is estimated based on a series of historical data. In cluster analysis, historical data are used to build a model that will measure the "distance" of a claim from the typical claims submitted by that provider or for that type of service. Decision trees use a series of screens or yes/no questions to determine the probability that a claim is valid.

The output of each of these methods is the probability of a claim's validity that is expressed as a score.

The claim score is typically structured so that it is directly related to the probability that a claim is in error. A high score may indicate a high probability that a claim is not legitimate. If the score meets a criteria (either above or below a cutoff value), then it is identified as a potential error.

The criteria or cutoff value may be used to tune the model to control the sensitivity and specificity of the model. If the cutoff is too extreme, then the model may not be sensitive enough and will allow fraudulent claims to be paid. If the cutoff is not extreme enough, then the model may not be specific enough and identify a large number of false positives.

In the healthcare setting, clinicians can use predictive modeling to improve patient care—one example would be to use predictive modeling to prevent readmission. For example, if you know the patient has a certain medical condition and lives alone, you might predict that they are a higher risk for readmission, but through prediction, the clinician can take preventive action and possibly mitigate the readmission risk.

6. White, Susan. "Predictive Modeling 101." *Journal of AHIMA* vol. 82, no. 9 (September 2011): 46–47.

FORMULAS AND STATISTICS

Statistical analysis typically is segmented into two areas: descriptive and inferential. Descriptive statistics, as the name would imply, are used to describe the characteristics of a set of data. They include measures of central tendency and measures of variation or dispersion. Inferential statistics are used to make decisions on the basis of sampled data, and they include measures such as confidence intervals (CIs) and hypothesis testing. Both are discussed in more detail in this section.

Descriptive Statistics

MEASURES OF CENTRAL TENDENCY

These statistics measure the center of a distribution. They are well suited for describing the typical value of a particular data element.

MEAN

Also known as:

Arithmetic average, arithmetic mean, average

Notation:

$\bar{x}$ or x-bar is the typical symbol used for the mean.

Properties:

The mean is the most common measure of central tendency. It is appropriate to use for continuous variables (charges, LOS, systolic blood pressure). The mean is not appropriate for use with nominal variables (categories in which order does not convey information, such as sex, race, or CPT codes). The mean can be influenced by outliers and may not be the best choice of statistic for a heavily skewed variable.

Calculation:

The mean is the sum of the observations divided by the number of observations

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where x_i is the i^{th} observation and n is the number of observations to be averaged. Example:

The average of the observations 2, 6, 9, 1 is $\bar{x} = \frac{(2+6+9+1)}{4} = 4.5$.

Excel Command:

=average(a1..a4), where a1..a4 refers to the data cells

MEDIAN

Also known as:

50th percentile, 2nd quartile, middle value

Notation:

There is no standard notation for the median. The symbol $\tilde{x}$ or x-tilde is sometimes used.

Properties:

The median is less influenced by outlier observations than is the mean.

Calculation:

To calculate the median of a series of observations, sort them from smallest to largest and choose the middle value as the median. If there is an even number of observations, then the median is the average of the two middle values.

Example:

To find the median of the observations 2, 6, 9, 1, first order the values (1, 2, 6, 9). There are an even number of observations (4); therefore, we average the two middle values to calculate the median:

$$\tilde{x} = \frac{(2+6)}{2} = 4$$

Excel Command:

=median(a1..a4) or =quartile(a1..a4,2), where a1..a4 refers to the data cells

Note that Excel and other software packages may each use different algorithms to estimate the median, so the results for an even numbered set of values may not match the hand-calculated median.

MODE

Also known as:

Value with the highest frequency or most "popular" value

Notation:

There is no standard notation for the mode.

Properties:

The mode is appropriate for use with categorical variables (sex, diagnosis-related group [DRG], etc.). The mode should not be used to describe the center of the distribution of a continuous variable (charges, time, etc.).

Calculation:

To calculate the mode of a set of observations, count the frequency of each value and select the value with the highest frequency. There may be multiple values of the mode if two or more values have the same (maximum) frequency. If there are two modes, the distribution is called "bimodal."

Example:

To find the mode of the observations 2, 6, 9, 1, 2, note that each value except 2 occurs once. The value 2 occurs twice and is, therefore, the mode.

Excel Command:

=mode(a1..a4), where a1..a4 refers to the data cells

PERCENTILES

Also known as:

25^{th} percentile = first quartile; 50^{th} percentile = second quartile or median; 75^{th} percentile = third quartile. The minimum and the maximum are special cases of the percentile.

Notation:

p_k, where k is the k^{th} percentile. For instance, $p_5 = 5^{th}$ percentile.

Properties:

Percentiles are most appropriate for continuous variables (ratio or interval). The minimum and maximum obviously are influenced heavily by outliers. Quartiles and midpercentiles are not heavily influenced by outliers. Quartiles often are used in benchmarking financial data.

Calculation:

To calculate percentiles of a distribution, you should first order the values. The k^{th} percentile is a value, p_k, such that most $(100k)$% of the measurements are less than this value and most $100(1 - k)$% are greater. First calculate $kN/100$, where k is the k^{th} percentile and N is the number of observations. If $kN/100$ is not an integer, then round up to the next integer. The k^{th} percentile is the k^{th} ordered observation. Computer programs such as Excel and SPSS may use different algorithms to estimate percentiles. Hand calculations likely will give a slightly different answer.

Example:

To find the 70^{th} percentile or p_{70} of the observations 2, 6, 9, 1, 2, first calculate $kN/100$ where k is the k^{th} percentile and N is the number of observations (70*5/100 = 3.5). Round up to an integer: 4. The k^{th} percentile is the k^{th} ordered observation; thus the 70^{th} percentile of 1, 2, 2, 6, 9 is $p_{70} = 6$.

Excel Command:

=percentile(a1..a4, k), where a1..a4 refers to the data cells and k is the percentile desired

GEOMETRIC MEAN

Also known as:

GM (as in GMLOS or geometric mean length of stay as reported by Centers for Medicare and Medicaid Services [CMS])

Notation:

p_k, where k is the k^{th} percentile. For instance, $p_5 = 5^{\text{th}}$ percentile.

Properties:

In the geometric mean, multiplication is used to summarize the variables. The geometric mean is appropriate to use for positive continuous variables. The geometric mean is always smaller than the arithmetic mean or average. The geometric mean is less influenced by large positive outliers than is the arithmetic mean; therefore, if a distribution has a long tail or is positively skewed, the geometric mean is a good measure of the center of the distribution. It is fairly common in the finance industry. CMS uses the geometric mean to summarize lengths of stay according to Medicare severity-adjusted DRG (MS-DRG). Lengths of stay are positive variables and tend to have a long tail (a few patients have very long stays).

Calculation:

The geometric mean is the n^{th} root of a series of n observations. The formula for the GM is $GM = \sqrt[n]{\prod_{i=1}^{n} x_i}$, where n is the number of observations, or equivalently $GM = \left(\prod_{i=1}^{n} x_i\right)^{1/n}$, where n is the number of observations. (The $1/n$ power and the n^{th} root are mathematically equivalent.)

Example:

To find the GM of the observations 2, 6, 9, 1, 2:

$$\sqrt[5]{2x6x9x1x2} = \sqrt[5]{216} = 2.93$$

Excel Command:

=geomean(a1..a4, k), where a1..a4 refers to the data cells and k is the percentile desired

Measures of Variation or Dispersion

SAMPLE OR POPULATION VARIANCE

Also known as:

Most reported variance values are based on a sample and therefore should be referred to as the "sample variance." In practice, the shorthand "variance" typically is used.

Notation:

Population variance: σ^2
Sample variance: s^2

Properties:

The population variance is the average of the squared deviations from the population mean. Since the values of any variable rarely are known for the entire population, the sample variance typically is used. The deviation (or difference) between each value and the mean is squared so that the "typical" deviation can be summarized regardless of the sign (or direction) of the deviation. The sum of the deviations from the sample mean is actually zero for any sample of observations, so the "squaring" allows the typical deviation to be quantified. The sample variance is somewhat influenced by outliers. The variance is an appropriate measure of spread for continuous variables (interval or ratio).

Calculation:

$$s^2 = \sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1},$$

where $\bar{x}$ is the sample mean and n is the number of observations.

Example:

To find the variance of the observations 2, 6, 9, 1, first calculate the sample mean, which is 4.5 (see above).

$$s^2 = \frac{(2-4.5)^2 + (6-4.5)^2 + (9-4.5)^2 + (1-4.5)^2}{4-1} = \frac{41}{3} = 13.67$$

Excel Command:

Population variance (denominator is n instead of $n - 1$):
=varp(a1..a4, k), where a1..a4 refers to the data cells

Sample variance:

=var(a1..a4, k), where a1..a4 refers to the data cells

STANDARD DEVIATION

Also known as:

Most reported standard deviation values are based on a sample and therefore should be referred to as the "sample standard deviation." In practice, the shorthand "standard deviation" typically is used.

Notation:

Population standard deviation: σ
Sample standard deviation: s

Properties:

The sample standard deviation is the square root of the sample variance. The units of measure of the variance is "squared units" (squared days, square inches, etc.) and is therefore less intuitive to apply than is the standard deviation. The standard deviation (sample or population) has the same unit of measure as the original observations. The sample standard deviation is somewhat influenced by outliers. The standard deviation is an appropriate measure of spread for continuous variables (interval or ratio).

Calculation:

$$s^2 = \sqrt{s^2}$$

where s^2 is the sample standard deviation.

Example:

To find the variance of the observations 2, 6, 9, 1, first calculate the sample mean, which is 4.5 (see above).

$$s = \sqrt{13.67} = 3.70$$

Excel Command:

Population standard deviation:
=stdevp(a1..a4, *k*), where a1..a4 refers to the data cells

Sample standard deviation:

=stdev(a1..a4, *k*), where a1..a4 refers to the data cells

RANGE

Also known as:

Range of values, spans

Notation:

There is no standard notation for the range.

Properties:

The range is the maximum (largest) value minus the minimum (smallest) value of a series of data points. The range is heavily influenced by outliers. The range is an appropriate measure of spread for most variable types (continuous and discrete). It is not appropriate for use with nominal data where order does not have a meaning (minimum and maximum would not have a meaning for that variable type either).

Calculation:

Range = maximum – minimum

Example:

To find the range of the observations 2, 6, 9, 1, first order the observations from smallest to largest: 1, 2, 6, 9. Note that the minimum value is 1 and the maximum value is 9. The range is, therefore, 9 – 1 = 8.

Excel Command:

There is no Excel command to calculate the range directly. You must use the min and max functions: =max(a1..a4, k) - min(a1..a4, k), where a1..a4 refers to the data cells.

STATISTICAL INFERENCE

"**Statistical inference** is the process of making conclusions using data that is subject to random variation, for example, observational errors or sampling variation."[7]

"More substantially, the terms **statistical inference, statistical induction** and **inferential statistics** are used to describe systems of procedures that can be used to draw conclusions from datasets arising from systems affected by random variation. Initial requirements of such a system of procedures for inference and induction are that the system should produce reasonable answers when applied to well-defined situations and that it should be general enough to be applied across a range of situations." [8]

These definitions cover most of the aspects of statistical inference, which is basically a set of statistical methods that is used to make decisions or conclusions on the basis of sample data.

That decision typically has a probability of error associated with it. The two most common types of statistical inference used in practice are confidence intervals and hypothesis tests.

7. Upton, Graham., and Ian Cook. *Oxford Dictionary of Statistics.* New York: Oxford University Press, 2008.
8. Dodge, Yadolah, ed. *The Oxford Dictionary of Statistical Terms.* New York: Oxford University Press, 2003.

Confidence Intervals

A confidence interval (CI) is a range of values that has a set probability or confidence level of containing the population value of the statistic of interest. CIs are encountered quite often in practice. Opinion polls are reported in the news with an associated +/– value. For instance, a 2010 Gallup poll reported that Democrats held an advantage over Republicans of 48 percent to 44 percent in a generic ballot poll. Each Gallup poll includes a section titled "Survey Methods." For this survey, Gallup reports:

> Results are based on telephone interviews conducted as part of Gallup Daily tracking survey July 19–25, 2010, with a random sample of 1,633 registered voters, aged 18 and older, living in all 50 U.S. states and the District of Columbia, selected using random-digit-dial sampling.
>
> For results based on the total sample of registered voters, one can say with 95% confidence that the maximum margin of sampling error is ±4 percentage points.[9]

In effect, Gallup is 95 percent certain that the percentage for Democrats is between 44 percent and 52 percent and the percentage favoring Republicans is between 40 percent and 48 percent. When stated in this manner, it is clear that this is a "statistical tie."

The CI is the probability of the interval covering the true value. In our example, Gallup set the confidence level at 95 percent for the percentage of voters favoring each party.

The width of the CI (the portion after the +/–) typically is referred to as the "precision." A narrower interval or smaller +/– value is more precise. The width of a CI depends on:

1. Sample size: increased sample size will result in increased precision.
2. Standard deviation of the variable: smaller standard deviation results in increased precision.
3. Decreased confidence level: decreased CI results in increased precision.

The formula to determine a CI depends on the statistic used as the estimator (percentage, arithmetic mean, etc.). The basic form of a CI is typically:

$$A(1-a)\%\ \textit{Confidence Interval} = (\textit{estimator}) \pm (\textit{critical value}) \times (\textit{standard error of estimator})$$

The critical value depends on the distribution of the estimator. If the arithmetic mean is used as the estimator, then the critical value is based on the normal or *t*-distribution. A larger confidence level typically results in a larger critical value and a wider interval.

More details on this topic can be found in most introductory statistics textbooks or via a quick Internet search.

9. Newport, Frank. "Democrats Maintain Advantage on Generic Ballot, 48% to 44%." Gallup Politics, July 26, 2010. http://www.gallup.com/poll/141557/Democrats-Maintain-Advantage-Generic-Ballot.aspx.

Hypothesis Testing

In hypothesis testing, the analyst is trying to make a decision, on the basis of sample data, between two hypotheses: the null hypothesis and the alternative hypothesis. The null hypothesis is typically the "status quo" and requires no action. The alternative hypothesis is sometimes called the "research hypothesis" and typically requires some action.

A test statistic is calculated to determine if the null hypothesis should be rejected or not on the basis of the evidence presented in the sample data, which is best illustrated with a simple example. Suppose the state average LOS for acute myocardial infarction is 4.2 days. The marketing department of Major Hospital would like to claim that their LOS is shorter than the state average. From a 10-patient sample during the last year, an average LOS of 3.5 days was calculated with a standard deviation of 1.1 days. A simple hypothesis test can help the marketing director understand if the data support that claim. The null hypothesis (Ho) is that the LOS at Major Hospital is at least as long as 4.2 days (state average), and the alternative hypothesis is that the LOS at Major Hospital is less than 4.2 days.

This situation calls for a *t*-test to be performed. After performing the *t*-test, compare the value of the test statistic to its distribution to determine the *P* value. The *P* value is the probability of obtaining a test statistic that large by chance. Alternatively, the *P* value is the probability of being wrong in concluding that the null hypothesis is false. In our example, the *P* value is the probability of being incorrect when claiming Major Hospital's LOS for patients with acute myocardial infarction is lower than the state average on the basis of the sample data presented.

The *P* value for this sample is 0.04, or 4 percent. Thus, if the marketing director is willing to take a 4 percent chance of making an incorrect conclusion, she can go forward and make the claim that Major Hospital's LOS is shorter.

Often an alpha level is set before performing a hypothesis test. The alpha level should be based on the cost of incorrectly rejecting the null hypothesis. In clinical studies, the alpha level is typically set low (1 percent or 5 percent). Since the *P* value is the probability of incorrectly rejecting the null hypothesis, the null hypothesis should be rejected if the *P* value is smaller than the alpha level.

More details on this topic can be found in most introductory statistics textbooks or via a quick Internet search.

SAMPLE JOB SKILLS AND RESPONSIBILITIES

This section discusses potential job skills and responsibilities required by a health data analyst at entry, middle, and senior levels. Please note, this is not an all-inclusive listing; rather, it is a guide for current professionals in these roles or those with an interest in further career opportunities.

ENTRY LEVEL

Education:

BA/BS (preferably in health information management, health services research, or health administration)

Experience:

Prior healthcare experience required. Previous health data analyst and report writing experience preferred.

Skills and Abilities:

- Has basic understanding of coding systems (CPT, DRG, ICD-9-CM, ICD-10-CM/PCS, National Drug Code [NDC])
- Demonstrates strong verbal and written communication skills
- Demonstrates excellent organizational and time management skills
- Exhibits keen attention to detail and problem-solving skills
- Knows Microsoft Office, especially Excel and Access

Preferred: RHIA, CHDA, master's degree (health administration, health informatics, or similar), knowledge of Structured Query Language (SQL)

Typical Job Description:

Responsibilities

Daily Operations

- Identify data problem areas and conduct research to determine best course of action
- Analyze and solve issues with legacy, current, and planned systems as they relate to the integration and management of patient data (e.g., review for accuracy in record merge and unmerge processes)
- Analyze reports of data duplicates or other errors to provide ongoing appropriate interdepartmental communication and monthly or daily data reports (e.g., related to the enterprise master patient index [EMPI])
- Monitor metadata for process improvement opportunities (e.g., monitoring orders for successful computerized physician order entry (CPOE) implementation)
- Identify, analyze, and interpret trends or patterns in complex data sets
- Monitor data dictionary statistics

Data Capture

- In collaboration with others, develop and maintain databases and data systems necessary for projects and department functions
- Acquire and abstract primary or secondary data from existing internal or external data sources
- In collaboration with others, develop and implement data collection systems and other strategies that optimize statistical efficiency and data quality
- Perform data entry, either manually or using scanning technology, when needed or required

Data Reporting

- In collaboration with others, interpret data and develop recommendations on the basis of findings
- Develop graphs, reports, and presentations of project results, trends, data mining
- Perform basic statistical analyses for projects and reports
- Create and present quality dashboards
- Generate routine and/or ad hoc reports

MID-LEVEL

Education:

Master's degree in health administration, public health, health informatics, or similar

Experience:

- Minimum of five years of experience in data management and analysis in a healthcare, managed care, or insurance setting required; strong data manipulation techniques and proficiencies required; analytical experience developing recommendations from claims data analysis
- Strong quality control ethic, ability to assess accuracy of output at report level, as well as investigate potential issues at a micro level
- Firm understanding of the nuances of the managed care business structure and the ability to apply this knowledge to analytical projects
- Strong problem solving and analytical skills

Skills and Abilities:

- Strong verbal and written communication skills with experience interacting with senior management
- Ability to present complex information in an understandable and compelling manner
- Project management experience
- Proficiency in Microsoft Word, Excel, Access, and PowerPoint
- Experience using SAS, SPSS, or other statistical package is desirable for analyzing large data sets
- Programming skills preferred, adept at queries and report writing, knowledge of SQL
- Knowledge of statistics, at least to the degree necessary to communicate easily with statisticians
- Experience in data mining techniques and procedures and knowing when their use is appropriate
- Knowledge of coding classification systems and terminologies
- Understanding of risk adjustment models
- Familiarity with outcome measurements
- Preferred: CHDA, project management certification, budgeting experience, understanding of database design

Typical Job Description:

- Work collaboratively with data and reporting and the database administrator to help produce effective production management and utilization management reports in support of performance management related to utilization, cost, and risk with the various health plan data; monitor data integrity and quality of reports on a monthly basis
- Work collaboratively with data and reporting in monitoring financial performance in each health plan
- Develop and maintain claims audit reporting and processes
- Develop and maintain contract models in support of contract negotiations with health plans
- Develop, implement, and enhance evaluation and measurement models for the quality, data and reporting, and data warehouse department programs, projects, and initiatives for maximum effectiveness
- Recommend improvements to processes, programs, and initiatives by using analytical skills and a variety of reporting tools
- Determine the most appropriate approach for internal and external report design, production, and distribution, specific to the relevant audience

SENIOR LEVEL

Education:

Master's degree in health administration, public health, health informatics, or similar

Experience:

- Minimum of five years of experience in healthcare analysis required, preferably within a large tertiary care hospital, academic medical center, or other large medical institution; background in quality improvement, health statistics, health services research, or healthcare outcomes research strongly preferred
- Two to four years of experience developing SQL or Procedural Language (PL)/SQL programs, preferably with Oracle Database, and developing reports by using Crystal Reports (Business Objects) or similar presentation tools
- Experience leading system-wide improvement projects in a matrix environment
- Minimum of three years of demonstrably successful managerial experience

Skills and Abilities:

- Strong verbal and written communication skills, with experience interacting with senior management and medical leadership
- Ability to present complex information in an understandable and compelling manner
- Understanding of healthcare finance and claims systems
- Ability to foster cooperation in a highly charged political environment
- Knowledge of No-SQL and/or Hadoop for large data sets
- Knowledge of SAS and SPSS
- CHDA preferred

Typical Job Description:

- Understand and address the information needs of governance, leadership, and staff to support continuous improvement of patient care processes and outcomes
- Lead and manage efforts to enhance the strategic use of data and analytic tools to improve clinical care processes and outcomes continuously
- Work to ensure the dissemination of accurate, reliable, timely, accessible, actionable information (data analysis) to help leaders and staff actively identify and address opportunities to improve patient care and related processes
- Work actively with information technology to select and/or develop tools to enable facility governance and leadership to monitor the progress of quality, patient safety, service, and related metrics continuously throughout the system
- Engage and collaborate with information technology and senior leadership to create and maintain a succinct report (e.g., dashboard), as well as a balanced set of system assessment measures, that conveys status and direction of key system-wide quality and patient safety initiatives for the trustee quality and safety committee and senior management; present this information regularly to the quality and safety committee of the board to ensure understanding of information contained therein
- Actively support the efforts of divisions, departments, programs, and clinical units to identify, obtain, and actively use quantitative information needed to support clinical quality monitoring and improvement activities
- Function as an advisor and technical resource regarding the use of data in clinical quality improvement activities
- Lead analysis of outcomes and resource utilization for specific patient populations as necessary
- Lead efforts to implement state-of-the-art quality improvement analytical tools (i.e., statistical process control)
- Play an active role, including leadership, where appropriate, on teams addressing system-wide clinical quality improvement opportunities

REFERENCE

AHIMA. "AHIMA Strategic Plan." 2013. http://library.ahima.org/xpedio/groups/public/documents/ahima/bok1_050165.pdf.

ADDITIONAL RESOURCES

Agency for Healthcare Research and Quality. "Data Sources Available from AHRQ." http://www.ahrq.gov/research/data/dataresources/.

Healthdata.gov. "CMS Medicare and Medicaid EHR Incentive Program, electronic health record products used for attestation." http://www.healthdata.gov/data/dataset/cms-medicare-and-medicaid-ehr-incentive-program-electronic-health-record-products-used.

National Center for the Analysis of Healthcare Data. "Map Samples." http://ncahd.org/map-samples.php.

GLOSSARY OF TERMS

CONTENTS

Calculations and Statistics 32
Classifications and Terminologies 34
Clinical Quality Measurement 35
Database Terms. 36
Finance and Reimbursement Terms 37
General 39

Calculations and Statistics

Alpha level: The probability of making a type I error. The alpha level should be based on the cost of incorrectly rejecting the null hypothesis. In clinical studies, the alpha level is typically set low (1 percent or 5 percent)

Analysis of variance (ANOVA): Test used to determine the differences among two or more means

Average length of stay (ALOS): The mean length of stay for hospital inpatients discharged during a given period

Bar chart: A graphic technique used to display frequency distribution data that fall into categories

Chi-square test: A statistical calculation used to determine whether proportions in a randomly drawn sample are significantly different from the underlying or theoretical population proportions

Confidence interval (CI): An interval that has a certain probability (confidence level) of including the true value of a population parameter. CIs may be calculated for population means, proportions, and standard deviations, etc.

Correlation: The existence and degree of linear relationship among factors

Dependent variable: A measurable variable in a research study that depends on an independent variable

Derived attribute: An attribute whose value is based on the value of other attributes (e.g., current date minus date of birth yields the derived attribute age)

Descriptive statistics: A set of statistical techniques used to describe data such as means, frequency distributions, and standard deviations; statistical information that describes the characteristics of a specific group or a population

***F* test:** The ratio of the between-group variance to the within-group variance in the ANOVA procedure. If the *F* ratio is statistically significant (the *F* value equals or exceeds the critical value of *F*), the observed differences between the group means of the independent variables under study will be significantly different from each other. See also **analysis of variance (ANOVA)**

Frequency distribution: A table or graph that displays the number of times (frequency) a particular observation occurs

Geometric mean length of stay (GMLOS): Statistically adjusted value of all cases of a given DRG, allowing for the outliers, transfer cases, and negative outlier cases that normally would skew the data. The GMLOS is used to compute hospital reimbursement for transfer cases

Hypothesis: A statement that describes a research question in measurable terms

Independent variable: An antecedent factor that researchers manipulate directly

Inferential statistics: Statistics that are used to draw conclusions regarding a population parameter on the basis of a sample

Length of stay (LOS): The total number of patient days for an inpatient episode, calculated by subtracting the date of admission from the date of discharge. If the admission and discharge are on the same day, the LOS generally is set to one day

Line graph: A graphic technique used to illustrate the relationship between continuous measurements; it consists of a line drawn to connect a series of points on an arithmetic scale and often is used to display time trends

Mean: A measure of central tendency that is determined by calculating the arithmetic average of the observations in a frequency distribution

Measures of central tendency: The typical or average numbers that are descriptive of the entire collection of data for a specific population

Median: A measure of central tendency that shows the midpoint of a frequency distribution when the observations have been arranged in order from lowest to highest

Mode: A measure of central tendency that consists of the most frequent observation in a frequency distribution

Normal distribution: A theoretical family of continuous frequency distributions characterized by a symmetric bell-shaped curve, with an equal mean, median, and mode; any standard deviation; and half of the observations above the mean and half below it

Null hypothesis: A hypothesis that states there is no association between the independent and dependent variables in a research study. The null hypothesis often represents the status quo or a state of no statistical difference in a study

Pie chart: A graphic technique in which the proportions of a category are displayed as portions of a circle (like pieces of a pie)

***P* value:** The probability that the observed difference could have been obtained by chance alone, given random variation and a single test of the null hypothesis

Qualitative analysis: In healthcare data, determining that the data accurately portray the care that was administered and that the content is correct

Quantitative analysis: In healthcare data, analyzing aggregate data for patterns. Typically, quantitative analysis requires numeric data for calculations

Range: A measure of variability that is the difference between the smallest and largest observations in a frequency distribution

Rank: Denotes a score's position in a group relative to other scores that have been organized in order of magnitude

Rate: A measure used to compare an event across time; it is a comparison of the number of times an event did happen (numerator) with the number of times an event could have happened (denominator). See also **ratio**

Ratio: A calculation performed by dividing one quantity by another. It is also a general term that can include a number of specific measures such as proportion, percentage, and rate

Regression analysis: Statistical technique that uses an independent variable to predict the value of a dependent variable. In the inpatient psychiatric facility prospective payment system (IPF PPS), patient demographics and length of stay (independent variables) were used to predict cost of care (dependent variable)

Sample: A set of elements drawn from and analyzed to estimate the characteristics of a population

Sample size: The number of subjects needed in a study to represent a population

Scatter diagram: A graph that visually displays relationships among factors

Standard deviation: A measure of variability that describes the deviation from the mean of a frequency distribution in the original units of measurement—the square root of the variance

Stratified random sampling: The process of selecting the same percentages of subjects for a study sample as they exist in the subgroups (strata) of the population

Systematic sampling: The process of selecting a sample of subjects for a study by drawing every n^{th} unit on a list

***t*-distribution:** The *t*-distribution is a probability distribution that is bell shaped and centered at zero, much like the standard normal distribution. The distribution is defined by the degrees of freedom, which determines the spread (or width) of the distribution. The *t*-distribution typically is used to determine the significance of a *t*-test of hypotheses regarding a population mean

***t*-test:** Assesses whether the means of two groups are statistically different from each other; appropriate when comparing the means of two groups

Trend: A long-term movement in an ordered series, say a time series, which may be regarded, together with the oscillation and random component, as generating the observed values[1]

Type I error: An error in which the researcher erroneously rejects the null hypothesis when it is true

Type II error: An error in which the researcher erroneously fails to reject the null hypothesis when it is false

Variable: A factor or quantity capable of assuming any of a set of values

Variance: A measure of variability that gives the average of the squared deviations from the mean—the difference between the planned or expected value and the actual value

1. Dodge, Yadolah, ed. *The Oxford Dictionary of Statistical Terms.* New York: Oxford University Press, 2003.

Classifications and Terminologies

Classification: A clinical vocabulary, terminology, or nomenclature that lists words or phrases with their meanings; provides for the proper use of clinical words as names or symbols; and facilitates mapping standardized terms to broader classifications for administrative, regulatory, oversight, and fiscal requirements

Current Procedural Terminology (CPT): A coding system used to report medical, surgical and diagnostic services provided in the outpatient and provider settings. Created and maintained by the American Medical Association.

Diagnosis-related group (DRG): A prospective payment mechanism for hospital inpatients in which diseases are placed into groups because related diseases and treatments tend to consume similar amounts of healthcare resources and incur similar amounts of cost; in the Medicare and Medicaid programs, one of the more than 700 diagnostic classifications in which cases demonstrate similar resource consumption and length-of-stay patterns. Prior to October 1, 2007, DRGs were not severity adjusted and were referred to as "CMS-DRGs." After October 1, 2007, DRGs are referred to as "Medicare severity-adjusted DRGs," or "MS-DRGs."

DRG grouper: A computer program that assigns inpatient cases to DRGs and determines the Medicare reimbursement rate

Evaluation and management (E/M) codes: Codes contained within CPT, developed by the American Medical Association, used for documenting physician services for billing purposes. These codes represent the level of service performed by the physician

Healthcare Common Procedure Coding System (HCPCS): A classification system to identify healthcare procedures, equipment, and supplies for claim submission purposes. There are two levels: level I CPT codes, developed by the American Medical Association, and level II codes for equipment, supplies, and services not covered by CPT codes, as well as modifiers that can be used with either level of codes, developed by CMS

International Classification of Diseases 9th Revision Clinical Modification (ICD-9-CM): Coding and classification system used to report diagnoses in all healthcare settings and inpatient procedures and services

International Classification of Diseases 10th Revision Clinical Modification (ICD-10-CM): Coding and classification system used to report diagnoses in all healthcare settings and inpatient procedures and services

Logical Observation Identifiers Names and Codes (LOINC): A database protocol aimed at standardizing laboratory and clinical codes for use in clinical care, outcomes management, and research. LOINC is currently required by CMS for identifying and reporting laboratory tests under Meaningful Use. Developed and maintained by the Regenstrief Institute for Health Care. Regenstrief and the International Health Terminology Standards Organisation (IHTSDO) entered into an agreement in 2013 to develop a linkage between LOINC and SNOMED

Major complication or comorbidity: A complication or comorbidity that increases the severity of a patient's condition and allows for assignment of an inpatient case to the highest-weighted DRG in the MS-DRG system

Major diagnostic category (MDC): Under DRGs, one of 25 categories based on single- or multiple-organ systems into which all diseases and disorders relating to that system are classified

Medicare severity-adjusted DRG (MS-DRG): See diagnosis-related group (DRG)

National Drug Code (NDC): Codes used to describe a drug on the basis of the manufacturer, the drug name, and the package size and type

RxNorm: Provides normalized names for clinical drugs and links those names to many of the drug vocabularies commonly used in pharmacy management and drug interaction software.These systems provide a link to the RxNorm identifier using the NDC. RxNorm is currently required by CMS for reporting pharmacy data under Meaningful Use. It is maintained by the National Library of Medicine (NLM)

Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT): A comprehensive clinical terminology, originally created by the College of American Pathologists and, as of April 2007, owned, maintained, and distributed by the International Health Terminology Standards Development Organisation (IHTSDO), a not-for-profit association in Denmark. The College of American Pathologists continues to support SNOMED CT operations under contract to the IHTSDO and provides SNOMED-related products and services as a licensee of the terminology

Therapeutic classification codes: A classification system that groups drugs into classes on the basis of their pharmacologic uses, such as antihypertensives and antibiotics. The AHFS Pharmacologic-Therapeutic Classification was developed and is maintained by the American Society of Health-System Pharmacists (ASHP)

Clinical Quality Measurement

Clinical Document Architecture (CDA): An HL7 document markup standard that specifies the structure and semantics of "clinical documents" for the purpose of exchange between healthcare providers and patients.

Clinical Quality Measure (CQM): Tools to help measure and track the quality of healthcare services provided by physicians, nurses, hospitals, and others in our health care system. CQMs are required as part of Meaningful Use requirements for the Medicare and Medicaid Electronic Health Record (EHR) incentive programs

Continuity of Care Document (CCD): A joint effort of HL7 International and ASTM to foster interoperability of clinical data by allowing physicians to send electronic medical information to other providers without loss of meaning and enabling improvement of patient care. The CCD establishes a rich set of templates representing the typical sections of a summary record, and expresses these templates as constraints on Clinical Document Architecture (CDA)

eMeasure: A health quality measure formatted using the HL7 Health Quality Measure Format (HQMF) standard specification

Health Quality Measure Format (HQMF): A standard for representing health quality measures as an electronic document. HQMF formally defines a quality measure (data elements, logic, definitions, etc.) to support consistent and unambiguous interpretation. Quality measure developers can encode their measures in this format so that they can be consumed by provider organizations, which will then be able to use the formal definitions to, for instance, query their EHR system

Measure Authoring Tool (MAT): A publicly available, web-based tool for measure developers to author electronic Clinical Quality Measures (eCQMs) using the Quality Data Model (QDM). Originally developed by the National Quality Forum under contract from CMS, as of January 2013, CMS assumed ownership of the tool

Quality data element: Data criteria within a QDM eMeasure composed of Quality Data Type, Quality Data Attribute, and Value Set

Quality Data Model (QDM): An "information model" that clearly defines concepts used in quality measures and clinical care and is intended to enable automation of structured data capture in electronic health record (EHR) and other electronic data sources. Developed and maintained by the National Quality Forum

Quality Reporting Document Architecture (QRDA): A document format that provides a standard structure with which to report quality measure data to organizations that will analyze and interpret the data. The Office of the National Coordinator for Health IT (ONC) adopted QRDA Category I (patient level) and QRDA Category III (aggregate) data submission approaches for Meaningful Use Stage 2 reporting.

QRDA Category I: individual patient-level report containing data for one or more quality measures

QRDA Category III: an aggregate level report containing calculated summary data for one or more measures

Value Set: One or more coded values used to filter patient populations for a specific data criterion (e.g., race)

Database Terms

Clinical data repository (CDR): A central database that focuses on clinical information

Data dictionary: A descriptive list of the data elements to be collected in an information system or database, the purpose of which is to ensure consistency of terminology

Data element: An individual fact or measurement that is the smallest unique subset of a database

Data mart: A well-organized, user-centered, searchable database system that usually draws information from a data warehouse to meet the specific needs of users

Data mining: The process of extracting information from a database and then quantifying and filtering discrete, structured data

Data repository: An open-structured database that is not dedicated to the software of any particular vendor or data supplier, in which data from diverse sources are stored so that an integrated, multidisciplinary view of the data can be achieved; also called a "central data repository" or, when related specifically to healthcare data, a "clinical data repository"

Data warehouse: A database that makes it possible to access information from multiple databases and combine the results into a single query and reporting interface. See also clinical data repository

Database: An organized collection of data, text, references, or pictures in a standardized format, typically stored in a computer system for multiple applications

Decision support system (DSS): A computer-based system that gathers data from a variety of sources and assists in providing structure to the data by using various analytical models and visual tools to facilitate and improve the ultimate outcome in decision-making tasks associated with nonroutine and nonrepetitive problems

Edit: A condition that must be satisfied before a computer system can accept data

Interface engine: A computer program that isolates the task of transferring data from one database to another

Metadata: Data about data that describe a specific item's content

Query: The process of making a logical inquiry or request from a database

Structured Query Language (SQL): A fourth-generation computer language that includes both data definition language and data manipulation language components and is used to create and manipulate relational databases

Finance and Reimbursement Terms

Actual charge: 1. A physician's actual fee for service at the time an insurance claim is submitted to an insurance company, a government payer, or a health maintenance organization; it may differ from the allowable charge 2. The amount a provider actually bills a patient, which may differ from the allowable charge

Allowable charge: Average or maximum amount a third-party payer will reimburse providers for a service

Ambulatory payment classification (APC) system: The prospective payment system used since 2000 for reimbursement of hospitals for outpatient services provided to Medicare and Medicaid beneficiaries

Case mix: A description of a patient population on the basis of any number of specific characteristics, including age, sex, type of insurance, diagnosis, risk factors, treatment received, and resources used

Case-mix index (CMI): The average relative weight of all cases treated at a given facility or by a given physician, which reflects the resource intensity or clinical severity of a specific group in relation to the other groups in the classification system; it is calculated by dividing the sum of the weights of DRGs for patients discharged during a given period by the total number of patients discharged

Claim: An itemized statement of healthcare services and their costs provided by the hospital, physician office, or other healthcare provider; it is submitted for reimbursement to the healthcare insurance plan by either the insured party or the provider

CMS-1500: A Medicare claim form used to bill third-party payers for provider services (e.g., physician office visits)

Cost report: A report that analyzes the direct and indirect costs of providing care to Medicare patients; it is required from providers annually for the Medicare program to make a proper determination of amounts payable to providers under its provisions

Explanation of benefits (EOB): A statement issued to the insured and the healthcare provider by an insurer to explain the services provided, amounts billed, and payments made by a health plan. See also **remittance advice (RA)**

Financial data: The data collected for the purpose of managing the assets of a business (e.g., a healthcare organization, a product line); in healthcare, data derived from the charge-generation documentation associated with the activities of care and then aggregated according to specific customer grouping for financial analysis

Geographic adjustment factor (GAF): Adjustment to the national standardized Medicare fee schedule relative value components used to account for differences in the cost of practicing medicine in different geographic areas of the country

Geographic practice cost index (GPCI): An index developed by CMS to measure the differences in resource costs among fee schedule areas compared to the national average in the three components of the relative value unit (RVU)—physician work, practice expenses, and malpractice coverage; a separate GPCI exists for each element of the RVU, and they are used to adjust the RVUs, which are national averages, to reflect local costs

Health plan: An entity that provides or pays for the cost of medical care, including a group health plan, a health insurance issuer, a health maintenance organization, or any welfare benefits plan such as Medicare, Medicaid, Civilian Health and Medical Program of Uniformed Services, and Indian Health Service

Hospital outpatient prospective payment system (HOPPS): See **outpatient prospective payment system (OPPS)**

Incurred but not reported (IBNR): One of the basic concepts of managing capitation contracts is estimating and accruing any IBNR expenses and liabilities

Inpatient prospective payment system (IPPS): A type of reimbursement system that is based on present payment levels rather than actual charges billed after the service has been provided; specifically, one of several Medicare reimbursement systems based on predetermined payment rates or periods and linked to the anticipated intensity of services delivered, as well as the beneficiary's condition

Medicare Administrative Contractor (MAC): An entity that processes all Medicare claims for both Part A and Part B by using the same common working file

Medicare code editor (MCE): A software program that detects and reports errors in the coding claims data submitted to the Medicare program

Medicare Provider Analysis and Review (MEDPAR) file: A file that contains data from Medicare claims for services provided to beneficiaries admitted to Medicare-certified inpatient hospitals and skilled nursing facilities

National Provider Identifier (NPI): A 10-digit number that uniquely identifies a healthcare provider

Outpatient code editor (OCE): A software program linked to the National Correct Coding Initiative that applies a set of logical rules to determine whether various combinations of codes are correct and appropriately represent the services provided

Outpatient prospective payment system (OPPS): The Medicare prospective payment system is used for hospital-based outpatient services and procedures and is predicated on the assignment of APCs

Outpatient service-mix index (OSMI): The sum of the weights of ambulatory payment classification groups for patients treated during a given period divided by the total volume of patients treated

Outpatient visit: A patient's visit to one or more units located in the ambulatory services area (clinic or physician's office) of an acute care hospital

Per member per month (PMPM): A common method to express healthcare costs on the basis of a single member in a month

Prospective payment system (PPS): A type of reimbursement system that is based on preset payment levels rather than actual charges billed after the service has been provided— specifically, one of several Medicare reimbursement systems based on predetermined payment rates or periods and linked to the anticipated intensity of services delivered, as well as the beneficiary's condition

Recovery audit contractor (RAC): Entity operating on behalf of CMS to identify improper payments made on claims of healthcare services provided to Medicare beneficiaries. Improper payments may be overpayments or underpayments and may be based on coding or medical necessity policies

Relative value unit (RVU): A measurement that represents the value of practice expense, malpractice expense, and physician work involved in providing a specific professional medical service in relation to the value of other medical services

Remittance advice (RA): Report sent by third-party payer that outlines claim rejections, denials, and payments to the facility. See also **explanation of benefits (EOB)**

Resource-Based Relative Value Scale (RBRVS): A Medicare reimbursement system to compensate physicians according to a fee schedule predicated on weights assigned on the basis of the resources required to provide the services

Revenue: The charges generated from providing healthcare services—earned and measurable income

Revenue code: A four-digit number in the chargemaster that totals all items and their charges for printing on the form used for Medicare billing

Revenue cycle: The process of how patient financial and health information moves into, through, and out of the healthcare facility, culminating with the facility receiving reimbursement for services provided—the regularly repeating set of events that produces revenue

Service-mix index (SMI): Describes the outpatient population in a single number; it is determined by multiplying the number of cases times the relative weight for the particular ambulatory payment classification (APC), summing the relative weights, and dividing by the total number of cases

Uniform Bill-04 (UB-04): The single standardized Medicare form for uniform billing, implemented in 2007 for hospital inpatients and outpatients; this form is used by the major third-party payers and most hospitals

General

Accuracy: A characteristic of data that are free from significant error and are up-to-date and representative of relevant facts

Administrative data: Data created or collected through the registration and billing process

Aggregate data: Data extracted from individual health records and combined to form de-identified information about groups of patients, which can be compared and analyzed

Benchmarking: An analysis process based on comparison; a comparison of performance against a standard point of excellence, either within the organization (e.g., from year to year) or among organizations on specified variables (for example, length of stay or cost per DRG)

Categorical data: Types of data (nominal, ordinal) that represent values or observations that can be sorted into a category

Clinical data: Data captured during the process of diagnosis and treatment

Coded data: Data that are translated into a standard nomenclature of classification so that they can be aggregated, analyzed, and compared

Coding: The process of assigning numeric or alphanumeric representations to clinical documentation

Comorbidity: Pre-existing condition that, because of its presence with a specific diagnosis, causes an increase in length of stay by at least one day in approximately 75% of the cases (as in complication and comorbidity)

Complication: A medical condition that arises during an inpatient hospitalization (e.g., a postoperative wound infection)

Data: The dates, numbers, images, symbols, letters, and words that represent basic facts and observations about people, processes, measurements, and conditions

Data analysis: The process of looking at and summarizing data with the intent to extract useful information and develop conclusions

Data capture: The process of recording healthcare-related data in a health record system or clinical database

Data comparability: The standardization of vocabulary such that the meaning of a single term is the same each time the term is used. Data comparability produces consistency of information derived from those data

Data integrity: The extent to which healthcare data are complete, accurate, consistent, and timely

Data set: A list of recommended data elements with uniform definitions that are relevant for a particular use

Data stewardship: Responsibilities and accountabilities associated with managing, collecting, viewing, storing, sharing, disclosing, or otherwise making use of personal health information

Demographic information: Information used to identify an individual, such as name, address, sex, and age

Diagnostic data: Data obtained when diagnoses or the reasons for visit are coded

Discrete data: Data that represent separate and distinct values or observations (i.e., data that contain only finite numbers and have only specified values)

Grouper: A computer software program that automatically assigns prospective payment groups on the basis of clinical codes

Health information management (HIM) professional: An individual who has received professional training at the associate or baccalaureate degree level in the management of health data and flow of information throughout the healthcare delivery system

Healthcare Cost and Utilization Project (HCUP): A group of healthcare databases and related software tools developed through collaboration by the federal government, state governments, and industry to create a national information resource for patient-level healthcare data

Healthcare Effectiveness Data and Information Set (HEDIS): A data set used by health plans to collect data about the quality of care and service they provide

Hospital-acquired condition (HAC): Condition that is not present at the time of inpatient admission, resulting in the assignment of the case to a DRG that has a higher payment when the condition is present as a secondary diagnosis and could reasonably have been prevented by following evidence-based guidelines

Inpatient: A patient who is provided with room, board, and continuous general nursing services in an area of an acute care facility where patients generally stay at least overnight

Leapfrog Group: Organization that promotes healthcare safety by giving consumers the information they need to make better-informed choices about the hospitals they choose

Map: The process of linking the content of one classification or vocabulary system to another

Medical Group Management Association (MGMA): A professional association for executives within physician practices

Outpatient: A patient who receives ambulatory care services in a hospital-based clinic or department

Present on admission (POA) indicator: A code that describes whether the patient's diagnosis was present at the time of admission to the hospital

Primary data source: A record developed by healthcare professionals in the process of providing patient care. See also **secondary data source**

Principal diagnosis: The reason established after study to be chiefly responsible for occasioning the admission of the patient to the hospital for care. See also **secondary diagnosis**

Procedural data: Data obtained when procedures are coded by using a procedural classification and nomenclature system

Secondary data source: Data derived from the primary patient record, such as the data contained in an index or a database. See also **primary data source**

Secondary diagnosis: A statement of conditions coexisting during a hospital episode that affects the treatment received or the length of stay. See also **principal diagnosis**

Table: An organized arrangement of data, usually in columns and rows

ANNOTATED BIBLIOGRAPHY

This is a listing of commonly used Web sites, online courses, publications, and other helpful resources that provide related learning. It is not meant to be all-inclusive.

CONTENTS

Calculations and Statistics....41
Classifications and Terminologies....42
Database Terms....43
Finance and Reimbursement Terms....44
General....45
Certification....47

Calculations and Statistics

Calculating and Reporting Healthcare Statistics, 4th Edition
By Loretta Horton, 2012. An excellent reference for any healthcare professional interested in acquiring knowledge and sharpening skills in this area. Explanations of why and how each kind of statistic is calculated and used are reinforced with exercises to practice compiling inpatient service days, average length of stay and occupancy, mortality rates, and other important data. Data presentation, inferential statistics, and basic research principles are also covered.

Free Statistical Calculator Software
This page contains links to free software packages that you can download and install on your computer for stand-alone computing. http://statpages.org/javasta2.html

Healthcare Data and SAS
By Marge Scerbo, 2001. A how-to guide for SAS analysts working with health plan data.

Pennsylvania Health Care Cost Containment Council (PHC4)
PHC4 was one of the first organizations to release risk-adjusted comparative healthcare statistics.
http://www.phc4.org/

SAS
SAS is a vendor of easy-to-use business analytics software and services.
http://www.sas.com/

SPSS
IBM SPSS Statistics is a comprehensive, easy-to-use set of predictive analytic tools for business users, analysts, and statistical programmers. http://www.spss.com/

Statistical Applications for Health Information Management
By Carole Osborn, 2006. Copublished with Jones & Bartlett, this book covers the basic biostatistics, descriptive statistics, and inferential statistics that are unique to HIM.

Classifications and Terminologies

AHIMA
AHIMA. "Data Standards, Data Quality, and Interoperability(Updated). Appendix A: Data Standards Resource." *Journal of AHIMA* 84, no. 11 (November-December 2013): Web extra. http://library.ahima.org/

American Medical Association
CPT codes http://www.ama-assn.org/

Centers for Medicare and Medicaid Services (CMS)
ICD-9-CM codes http://www.cms.hhs.gov/ICD9ProviderDiagnosticCodes/01_overview.asp
ICD-10-CM codes http://www.cms.gov/ICD10/

Healthcare Code Sets, Clinical Terminologies, and Classification Systems
By Kathy Giannangelo, 2010. Describes the latest developments in the growing field of health informatics and multiple terminologies, vocabularies, code sets, and classification systems.

Healthcare Common Procedure Coding System (HCPCS) codes
http://www.cms.gov/HCPCSReleaseCodeSets/

International Health Terminology Standards Development Organisation (IHTSDO)
Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT)
http://www.ihtsdo.org/snomed-ct/

LOINC
LOINC laboratory codes
http://loinc.org/

National Drug Code Database
http://www.accessdata.fda.gov/scripts/cder/ndc/default.cfm

Therapeutic Classification Codes
http://www.ahfsdruginformation.com/pt-classification-system.aspx

Database Terms

ActiveReports
ActiveReports is a popular report writer. This site has good tips on writing effective reports.
http://www.componentone.com/SuperProducts/ActiveReports

"Beginner's Guide to Databases"
By Charles Nadeau. This article takes a case study through the database design from the very beginning—not a healthcare application, but a good discussion of the basics.
http://www.adobe.com/devnet/dreamweaver/articles/beginners_databases.html

"Database Design Basics"
This article gives an excellent overview of database design considerations. It is biased toward the features and functionality of Microsoft Access, but most of the concepts transfer to any database tool.
http://office.microsoft.com/en-us/access/HA012242471033.aspx

"Database Design from the Ground Up"
By Mike Chapple. This article is more advanced and discusses topics beyond the scope of this course. There are a number of good links to secondary articles.
http://databases.about.com/cs/specificproducts/a/designmenu.htm

"Data Dictionary"
This Wikipedia reference has a good description of a data dictionary and excellent references.
http://en.wikipedia.org/wiki/Data_dictionary

"Difference Between Reliability and Validity"
http://www.differencebetween.net/miscellaneous/difference-between-reliability-and-validity/

"Entity-Relationship Diagram"
By Mike Chapple. This article includes examples of entity-relationship diagrams.
http://databases.about.com/cs/specificproducts/g/er.htm

"Guidelines for Developing a Data Dictionary"
By the AHIMA e-HIM Work Group on EHR Data Content, *Journal of AHIMA* 77, no. 2 (February 2006): 64A–D.
http://library.ahima.org/

IBM Dictionary of Computing, 10th Edition
This reference includes technical definitions of database documentation.
http://portal.acm.org/citation.cfm?id=541721

"Reliability & Validity"
http://www.socialresearchmethods.net/kb/relandval.php

SAP Crystal Solutions
The SAP site gives more information about the product Crystal Reports. It is a very common report program that is used frequently with SQL databases.
http://www.sap.com/solutions/sapbusinessobjects/sme/reporting/crystalreports/index.epx

"Types of Reliability"
This article includes a good summary of the various types of reliability. It includes examples as well as theoretical descriptions of the concepts.
http://changingminds.org/explanations/research/design/types_reliability.htm

"What is the difference between data integrity and data validity?"
This Wiki.answers.com resource gives an excellent description of the difference between data validity and data integrity.
http://wiki.answers.com/Q/What_is_the_difference_between_data_integrity_and_data_validity

Finance and Reimbursement Terms

"Apple to Apples — RVU Analysis in Radiology"
By Mindy Goldsmith, *Radiology Today* 6, No. 11 (May 30, 2005): 14.
http://www.radiologytoday.net/archive/rt_053005p14.shtml

Principles of Healthcare Reimbursement, 4th Edition
By Anne B. Casto and Elizabeth Forrestal, 2013. This book explains how reimbursement systems affect providers and payers, consumers, policy makers, and the development of classification and information technology systems.

CMS Fact Sheets
CMS offers instructional guides explaining details of the forms and corresponding data element requirements.

CMS-1500
http://www.cms.gov/MLNProducts/downloads/form_cms-1500_fact_sheet.pdf

Medicare Claims Processing Manual (CMS-1450 Data Set)
http://www.cms.gov/manuals/downloads/clm104c25.pdf

Medicare Claims Processing Manual (CMS-1500 Data Set)
http://www.cms.gov/manuals/downloads/clm104c26.pdf

UB-04
http://www.cms.gov/Outreach-and-Education/Medicare-Learning-Network-MLN/MLNProducts/Downloads/837I-FormCMS-1450-ICN006926.pdf

General

A Practical Approach to Analyzing Healthcare Data, 2nd Edition
By Susan White, 2013. Published by AHIMA, this book offers guidance to healthcare professionals and health information management (HIM) students on how to best analyze, categorize, and manage healthcare data.

Agency for Healthcare Research and Quality (AHRQ)
This federal agency conducts many research and quality studies throughout the United States and publishes its results, as well as provide resources to both clinical and nonclinical healthcare professionals. Much of what AHRQ does pertains to electronic systems, EHRs, and DSSs. http://www.ahrq.gov/

AHIMA Engage
Engage Communities consist of both members-only and public communities arranged under several specifically defined healthcare and health information management (HIM) domains. The communities contain strategically aligned content and forums focused on areas of importance to HIM professionals. These domains include:

- Coding, Classification, and Reimbursement
- Confidentiality, Privacy, and Security
- Consumer Engagement and Personal Health Information
- Health Informatics
- Health Information Technologies and Processes
- Healthcare Leadership and Innovation
- Information Governance and Standards

Topics related to this toolkit can be found in many of these domains.

Bridges to Excellence (BTE)
In partnership with Leapfrog, the BTE programs recognize and reward practices and clinicians who meet evidence-based performance measures.
http://www.hci3.org/what_is_bte

Centers for Disease Control and Prevention (CDC) WONDER
CDC Wonder provides a single point of access to a wide variety of public health data reports and systems.
http://wonder.cdc.gov/

CMS Data Navigator
http://dnav.cms.gov/

CMS Measures Management System
CMS developed a standardized system for developing and maintaining the quality measures used in its various accountability initiatives and programs. Known as the Measures Management System, CMS-funded measure developers (or contractors) should follow this core set of business processes and decision criteria when developing, implementing, and maintaining quality measures.
https://www.cms.gov/Medicare/Quality-Initiatives-Patient-Assessment-Instruments/MMS/index.html?redirect=/MMS/19_MeasuresManagementSystemBlueprint.asp

Food and Drug Administration, NDC Directory
The National Drug Code Directory is a universal product identifier for human drugs.
http://www.fda.gov/Drugs/InformationOnDrugs/ucm142438.htm

"From Figures to Facts: Data Quality Managers Emerge as Knowledge Leaders"
By Ruth Carol, *Journal of AHIMA* 73, no.10 (Nov./Dec. 2002): 24–28. http://library.ahima.org/

Healthcare Cost and Utilization Project (HCUP)
A project sponsored by AHRQ that provides a variety of healthcare databases and related software tools, products, and statistical reports to inform policy makers, health system leaders, researchers, and the public. http://www.hcup-us.ahrq.gov/

HealthGrades
HealthGrades is a consumer Web site providing reports and ratings for physicians, hospitals, and nursing homes. http://www.healthgrades.com/

Hospital Compare
Hospital Compare is a consumer-oriented Web site that provides information on how well hospitals provide recommended care to their patients. http://www.hospitalcompare.hhs.gov/

The Leapfrog Group
The Leapfrog Group is a voluntary program aimed at mobilizing employer purchasing power to alert America's health industry that big leaps in healthcare safety, quality, and customer value will be recognized and rewarded. This voluntary program posts results on many criteria for consumers and healthcare professionals to use in decision making and benchmarking. http://www.leapfroggroup.org

Medical Group Management Association
Data for physician practices is available in the MGMA Store, item #6429. www.mgma.com

National Association for Healthcare Quality (NAHQ)
NAHQ is a national professional organization dedicated to offering resources pertaining to quality improvement to healthcare professionals. http://www.nahq.org/

Research Data Assistance Center
ResDAC is a government contractor that is charged with supporting the CMS public use data files. The Web site contains overviews of all files available for purchase from CMS, as well as detailed instructions regarding how to purchase data. http://www.resdac.umn.edu/

Value Set Authority Center (VSAC)
The Value Set Authority Center (VSAC) is provided by the National Library of Medicine (NLM), in collaboration with the Office of the National Coordinator for Health Information Technology and the Centers for Medicare & Medicaid Services, VSAC currently serves as the authority and central repository for the official versions of value sets that support Meaningful Use 2014 Clinical Quality Measures (CQMs). The VSAC provides search, retrieval, and download capabilities through a Web interface and APIs. https://vsac.nlm.nih.gov/

Certification

Certified Health Data Analyst (CHDA) Credential from AHIMA
This Web site provides an overview of the credential, including eligibility, examination information, and other frequently asked questions. http://www.ahima.org/certification/chda.aspx

CHDA Exam Prep Series
This is a three-part AHIMA distance education series designed to help students prepare for the CHDA certification examination. The three courses in the program are described below and can be purchased at http://www.ahima.org/education/onlineed/Programs/examprep.

Exam Prep: CHDA Domain 1—Data Management
This six-lesson course will educate students primarily in data management, specifically in regard to data structures and architecture. Data models, in addition to maintenance of databases, will be addressed.

Exam Prep: CHDA Domain 2—Data Analytics
This six-lesson course will educate students primarily in data analysis. A review of qualitative and quantitative analysis and their importance to valid data analysis will be reviewed. Various organizational processes may change on the basis of the analyzed results. This course will identify specific examples that may be affected.

Exam Prep: CHDA Domain 3—Data Reporting
This six-lesson course will educate students primarily in data reporting. Once data have been analyzed, it is vital to present the results to the business owners of the data elements. This course will review the best practices to accomplish this, in addition to identifying potential organizational effects of the reported data.

CASE STUDY

Throughout the toolkit, guidance has been provided for how to solicit information needs from the requester appropriately, how to validate the data compiled, and how to display data. As an additional illustration, a published infection control and hospital epidemiology study has been included here to illustrate the path data take in a study from the point of collection to displaying the outcomes of the analysis.

Appendix C shows the study data definitions document used to collect data for the study, and the article referenced below discusses the output:

Lipsky, Benjamin, et al. "Skin, Soft Tissue, Bone, and Joint Infections in Hospitalized Patients: Epidemiology and Microbiological, Clinical, and Economic Outcomes." *Infection Control and Hospital Epidemiology* 28, no. 11 (November 2007). http://www.jstor.org/stable/10.1086/520743

APPENDIX A: Data Dictionary Sample

DATA FIELD	NAME	DEFINITION	DATA TYPE	FORMAT	FIELD SIZE	VALUES	SOURCE SYSTEM	DATE FIRST ENTERED	WHY ITEM IS INCLUDED
Admission Date	ADMIT_DATE	The date the patient is admitted to the facility as an inpatient	date	mmddyyyy	8	Admission date cannot precede birth date or 2007 No hyphens or slashes	Patient Census	2/23/2008	Allows analysis of patients and services within a specific period that can be compared with other periods or trended
Census	CENSUS	The number of inpatients present in the facility at any given time	numeric	x to xx	3	Any whole number from 0 to 999	Patient Census	2/23/2008	Provides analysis of budget variances, aids future budgetary decisions, and allows quicker response to negative trends
Ethnicity	PT_ETHNIC	Patient's ethnicity Must be reported according to official Office of Management and Budget categories	alphanumeric	Ex; letter must be uppercase	2	E1 = Hispanic or Latino Ethnicity E2 = Non- Hispanic or Latino Ethnicity	Patient Census; Practice Management	2/23/2008	Patient demographics aid marketing and planning future budgets and services
Infant Patient	INFANT_PT	A patient who has not reached 1 year of age at the time of discharge	alphanumeric	Age in months = xD to xxD OR xM to xxM	3	Must be > 0 AND < 1 year	Patient Census; Practice Management	2/23/2008	Patient age affects types of services required and payer sources
Inpatient Daily Census	IP_DAY_CENSUS	The number of inpatients present at census-taking time each day, plus any inpatients who were both admitted and discharged after the previous day's census-taking time	numeric	x to xx	3	Any whole number from 0 to 999	Patient Census	2/23/2008	Provides analysis of budget variances, aids future budgetary decisions, and allows quicker response to negative trends
Medical Record Number	MR_NUM	The unique number assigned to a patient's medical record The medical record is filed under this number	alphanumeric	xxxxxx: requires leading zeros	6	000001 to 999999	Patient Census; Practice Management		Provides analysis of services, resource utilization, and patient outcomes at the physician level
Patient Age	PT_AGE	Age of patient calculated by using most recent birthday attained before or on same day as discharge	numeric or alphanumeric	Age in days = xD to xxD OR Age in months = xM to xxM OR Age in years = x to xxx	3	Age must be > 0, and < OR = 124 years; children less than 1 year must be > 0 M AND < 1 year	Patient Census; Practice Management	2/23/2008	Patient age impacts the services utilized and payer sources
Patient Sex	PT_SEX	Patient sex	alphanumeric	letter; must be uppercase	1	M = Male F = Female U = Unknown	Patient Census; Practice Management	2/23/2008	Patient sex impacts the services and specialties utilized
Patient Zip Code	PT_ZIP_CODE	Zip code of patient's residence	alphanumeric	xxxxx-xxxx	11	00000 to 99999; 00000 = Unknown 99999 = Foreign	Patient Census; Practice Management	2/23/2008	Patient demographics aid marketing and planning future budgets/services
Pediatric Patient	PED_PT	A patient who has not reached 18 years of age at the time of discharge	numeric	Age in days = xD to xxD OR Age in months = xM to xxM OR Age in years = x to xxx	3	Age must be > 0 AND < 18 years; children less than 1 year must be > 0 M AND < 1 year	Patient Census; Practice Management	2/23/2008	Patient age impacts the services utilized and payer sources

APPENDIX B: Study Report and Request Form

There are many factors to be considered when a client or customer asks for health data. For the analyst to gather, analyze, and report the data correctly, it is essential to understand the "who, what, where, when, and why" of the request.

One of the first questions typically asked of the requester is "What questions are you trying to answer or what do you specifically want to know?"

Understanding the question is not only essential to providing accurate and correct information but also helpful in determining the project's scope.

Although the following questions are not necessarily a complete list, getting the answers to them will help you determine how complex a project is and should provide you with enough information to pull the correct data.

1. **Which patients are to be studied?**
    a. **Time Period:** calendar year (CY)/federal fiscal year (FFY)

       Examples: Jan-Dec = CY Oct03-Sep04 = FFY04

    b. **Patient Type:**

       Examples: inpatient/outpatient

       If outpatient, specify which settings are to be included: hospice, home health, hospital outpatient, physician office, freestanding ambulatory surgery center, independent laboratory

    c. **Type of Service:**

       Examples: physician/supplier/facility

       Exclude/include: independent laboratories, freestanding ambulatory care facilities, etc.
       Exclude/include: nurses, chiropractors, physician assistants, psychologists

    d. **Age of Population:**

       Examples: children, working adults (i.e., 18–65), seniors (65+)

    e. **Financial Class:** Medicare only, commercial insurance only (Blue Cross and Blue Shield, Aetna, WellPoint, etc.), all payers

    f. **Data Criteria to be *included* in the study:**
        1. DRG, APC
        2. Diagnosis or procedure: principal, secondary
        3. Drug: specific route or age stratification

    g. **Data Criteria to be *excluded* from the study:**

    h. **Be clear in the use of AND and OR**

       Must two or three things all be true to qualify a patient for the study (a and b and c), or can any one of them be true (a or b or c)?

    i. **Are there subsets of patients who need to be considered?**

2. **What does the requester want to know about these patients?**
   a. Interest in counts of visits, patients, admissions, procedures?
   b. Is there a need for comparisons or trending of patient groups, time periods, or procedures?
   c. Are column and row labels needed (i.e., percentage, average, counts)?
   d. Where is the cutoff for reporting—at least 5 percent?
   e. Is there a need for trimming of outliers—yes or no?
   f. Is there a need for a number of separate reports?
   g. What are the summarization levels?
      - Three-digit versus five-digit code levels
      - Group breakouts/roll-ups
3. **In what format does the requester want to see the information? (example criteria shown)**
   a. Output: hard-copy report/electronic media/both
   b. File format: Excel/Access/PDF/text file/graphic layout (chart, graph, or table)
4. **What are the requester's timeline or turnaround requirements?**
5. **What is the requester's price range or budget, if known?**

APPENDIX C: Sample Data Definitions

SKIN/SOFT TISSUE BONE JOINT INFECTION (SSTBJ) DATA DEFINITIONS EXAMPLE

Project Name:
Skin, Soft Tissue, Bone, Joint Infections

Deadline:
Interim deliverables – MM/DD/YY
Final analysis- MM/DD/YY
Abstract/Poster Submission Deadline - MM/DD/YY
Manuscript Deadline = MM/DD/YY

Patient Population:
Acute Care encounters having age at admission >17 years with a principal diagnosis in the categories of osteo/septic arthritis, surgical site infections, cellulitis or other skin or soft tissue related infection or diabetes with a secondary diagnosis in one of the SSTBJ categories. Cases in MDC 14 and 15 (maternal and neonatal DRGs) are excluded from this study. Analysis only being done for those encounters having a positive culture grow out from one of the culture sites defined as relevant to the SSTBJ diagnosis and grouping.

Timeframe:
Discharges from MM/DD/YY - MM/DD/YY

Eligible Hospitals:
Hospitals must meet all of the following clinical collection criteria to be included:

- Minimum 85% of eligible cases having at least one clinical laboratory result and at least one vital sign during the admission period (one day prior through one day after admission date). This rate should also not go below 80% rate in any ½ calendar year period.
- Minimum 30% of eligible cases having at least one culture drawn and no less than 25% in any ½ calendar year period.

Hospital must have fairly stable volume of eligible cases based on a review of calendar year eligible case counts.

Hospital must show some evidence of collecting MRSA throughout the study period.

Format:
Tables will be filled in according to the following specifications. Data source will be the SSTBJ data mart with extracted and derived data in SAS for subsequent analysis and manipulation.

Deliverable:
Statistical analysis; creation of abstract/poster material, potential manuscript

SSTBJ Group Criteria:
Each eligible case will be assigned to one of the SSTBJ groups based on a combination of principal diagnosis, secondary diagnosis (in the case of Diabetes) and any appropriate culture recorded during admission period. Since principal diagnosis is being used and each SSTBJ group contains a mutually exclusive set of diagnosis codes, there is no need to apply any type of hierarchy in order to assign each case to a single group.

Cases will be evaluated for grouping criteria on 2 levels—first, at the culture drawn level, and then at the positive culture level. Since data presentation is at the positive culture level, the following grouping criteria are described on the positive culture level. Culture positive means positive for ANY organism recorded.

Positive Culture Group Criteria:

Group 1 - Osteo/Septic Arthritis

Principal diagnosis code from Group 1 list AND culture positive from Joint/Muscle, Bone, Extremity, Skin culture sources during the admission period.

Group 2 – Surgical Wound

Principal diagnosis code from Group 2 list (surgical wound or Device/Prosthesis infection) AND culture positive from Joint/Muscle, Bone, Extremity, Skin or Device/Prosthesis culture sources during admission period. Note this group will be split into Device/Prosthesis infection and Other Surgical Wound.

Group 3 - Cellulitis

Principal diagnosis code from Group 3 list AND culture positive from Extremity or Skin culture sources during admission period.

Group 4 - Other SSTBJ Infections

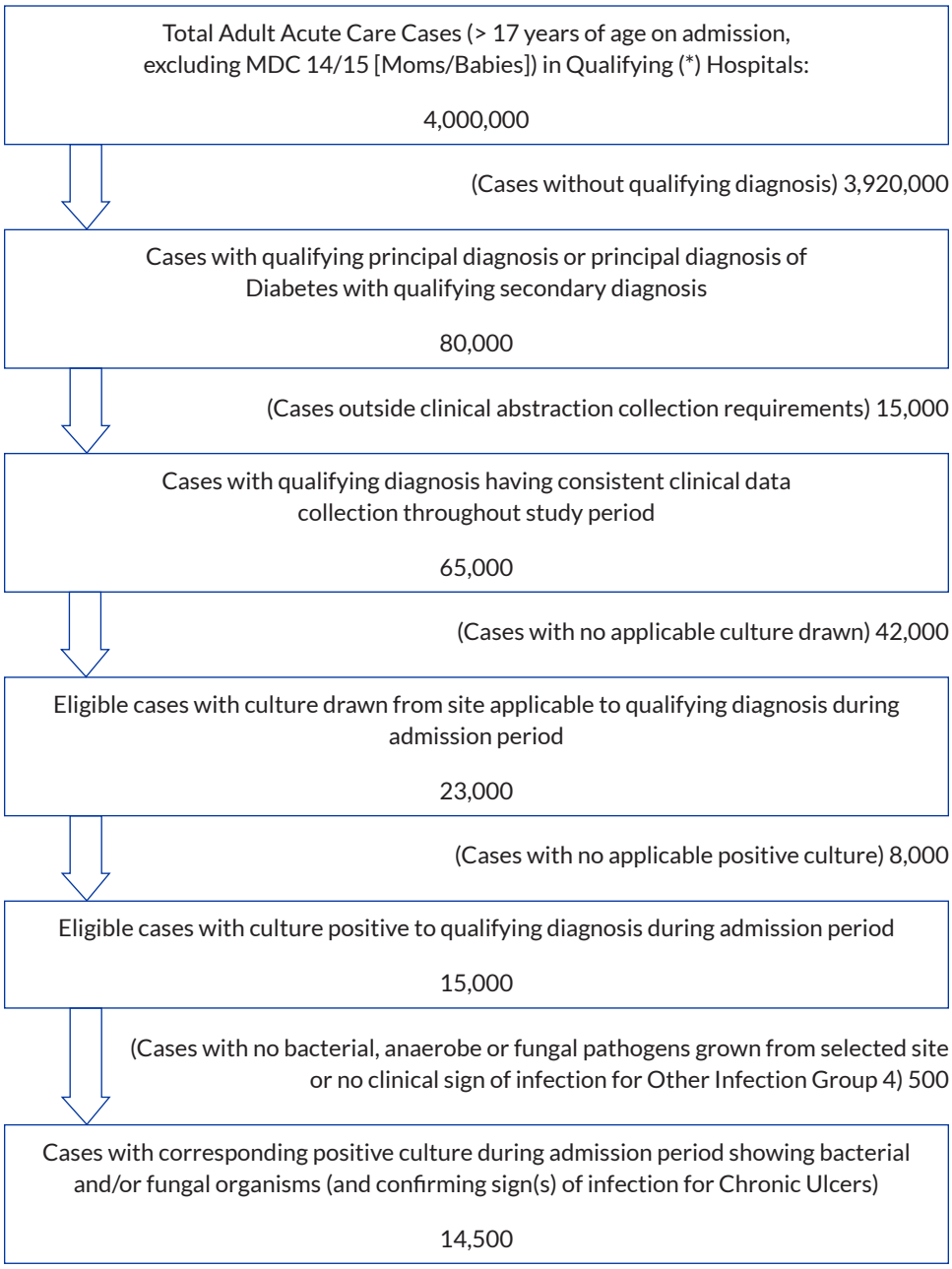
Principal diagnosis code for SSTBJ infection (group 4 list) not specified above AND culture positive from Joint/Muscle, Extremity or Skin culture sources during the admission period OR Principal diagnosis code of chronic ulcer AND culture positive from Joint/Muscle, Extremity or Skin culture sources and patient has WBC count >12,000 cells/mm3 or >5% neutrophilic bands or temperature of >38C or <36C during the admission period.

General Analysis Plan:

- Create together normalized data sets for:
  - Case characteristics (total charge, LOS, mortality)
    - ✓ Gender
    - ✓ Age
    - ✓ Transfer from SNF
    - ✓ Transfer from another acute care facility (healthcare associated vs. community)
    - ✓ Severity Risk Score
    - ✓ History of amputation
    - ✓ Disposition
    - ✓ ICU days
  - Case Outcomes
    - ✓ Total charge
    - ✓ LOS
    - ✓ Mortality
    - ✓ ICU days
  - Hospital characteristics: bed size, state, region, rural/urban
  - Eligible cultures for admission period
  - Lab/vital sign of interest for admission period
  - Procedures during admission period
  - Diagnosis codes for eligible encounters
  - Comorbid conditions for eligible encounters

PROJECT CASE TREE

The following is an example of creating a case tree that shows at which step and for what reason cases were removed from the final analysis population. This is an example of how a case tree would be created and does not reflect an actual analysis population.



(*) XX Hospitals qualified for this study based on the clinical data collection criteria itemized in the Eligible Hospitals Section

DEMOGRAPHICS AND CLINICAL CHARACTERISTICS LAYOUT AND DATA DEFINITIONS

Table 1 Demographics and clinical characteristics of SSTBJ patients

	GROUP 1: (n=)	GROUP 2: (n=)	GROUP 3: (n=)	GROUP 4: (n=)	TOTAL: (n=)
Race (n,%)					
Black					
White					
Other					
Unknown					
Not specified					
Ethnicity (n,%)					
Hispanic					
Non Hispanic					
Unknown					
Not specified					
Gender (n,%)					
Male					
Age (years) median, IQR, Mean, SD					
Insurance					
Managed care					
Medicaid					
Medicare					
Other					
Commercial					
Self pay					
Unknown					
Admitted from SNF (n,%)					
Acute Care Hospitalization in prior 12 months (n,%)					
ICU days (median, IQR, mean, SD)					
Ventilator support (n,%)					
Ventilator Patients					
Patients with > 95 Ventilator hours of those patients with ventilator support					
Surgery (a)					

(a) Surgery defined procedures performed under general, spinal anesthesia codes during the admission period

Data Definitions for Table 1

1. General Information
   a. Percents will be displayed to 1 decimal
   b. Unless otherwise specified, percentages are based on the total number of patients in the group column
   c. Bolded rows represent titles. No data will be displayed in that row
2. SSTBJ Group Columns

   Osteo/Septic Arthritis - number of patients in group 1
   Surgical Wound - number of patients in group 2
   Cellulitis - number of patients in group 3
   Other - number of patients in group 4
   Total - sum of all patients across the groups
3. Data Rows (includes data field names and corresponding coded values)
   a. Race (pt. race)

   Black - race = 40
   White - race = 50
   Other - race = 1, 20, 80
   Unknown - race = 99
   Not Specified - race = 0

   b. Ethnicity (pt. ethnicity)

   Hispanic - ethnicity = 1
   Non Hispanic - ethnicity = 6
   Unknown - ethnicity = 9
   Not Specified - ethnicity = 0

   c. Gender (pt. gender)

   Males - gender = 1

   d. Age (pt_sum. age)

   Median and lnterquartile Range for patients with age >17 years of age

   e. Insurance (pt.finclass)

   Managed care - finclass = 20,30
   Medicaid - finclass = 55
   Medicare - finclass = 50,53,22
   Other - finclass = 16-18, 57-85
   Commercial - finclass = 10,40
   Self pay - finclass = 1
   Unknown - finclass = 90

   f. Admitted from SNF (pt.admsrcer)

   admsrce = 18,41

   g. Acute Care Hospitalization in Prior 12 months (prior_ad.rec = 'AC' and prior_days <=365)

   h. ICU Days

   Sum of all careunit.cu_days for the admission where careunit between 300–380. Note for any calculation requiring the number of patients, only use those patients who had a stay in one of these careunits.

   i. Ventilator Patients

   Count of all patients with a px.code between 96.70–96.72 of all patients in the SSTBJ group column.

   j. Ventilator Patients on Ventilator more than 95 hours

   Count of all patients with a px.code of 96.72 of all patients with code between 96.70–96.72 for the patients in the SSTBJ group column.

   k. Surgery

   Count of patients with episode.anes_type = 25,41,45 and episode.day = –1, 0 or 1 for all patients in the SSTBJ group column.

Table 2a Organism distribution of SSTBJ groups for mono-microbial infections

MONO-MICROBIAL	GROUP 1: (n=)	GROUP 2: (n=)	GROUP 3: (n=)	GROUP 4: (n=)	TOTAL (n=)
Cultures drawn (n,%)					
Culture positive of cultures done in (n,%)					
I. AEROBES (TOTAL)					
A. Gram-Positives (All)					
1. S. aureus					
a. MSSA					
b. MRSA					
All MRSA as % of All S. aureus					
2. Coagulase-Negative Staphylococci*					
3. Streptococci					
a. Enterococci, Strep Group D (% VRE)					
b. Other Streptococci					
4. Other					
B. Gram-negatives (All)					
1. Enterobacteriaceae					
2. Pseudomonas spp.					
3. Other					
II. OBLIGATE ANAEROBES (TOTAL)					
A. Bacteroides spp.					
B. Clostridium spp.					
C. Other					
III. FUNGI (TOTAL)					
A. Candida spp.					
B. Other					
IV. VIRUSES (TOTAL)					
V. Other (TOTAL)					

*S. epidermis, Micrococci spp. & others; excludes coagulase-positive staphylococci

Table 2b Organism distribution by SSTBJ groups for poly-microbial infections

POLY-MICROBIAL	GROUP 1: (n=)	GROUP 2: (n=)	GROUP 3: (n=)	GROUP 4: (n=)	TOTAL (n=)
Cultures drawn (n,%)					
Culture positive of cultures done in (n,%)					
I. AEROBES (TOTAL)					
A. Gram-Positives (All)					
1. S. aureus					
a. MSSA					
b. MRSA					
All MRSA as % of All S. aureus					
2. Coagulase-Negative Staphylococci**					
3. Streptococci					
a. Enterococci, Strep Group D (% VRE)					
b. Other Streptococci					
4. Other					
B. Gram-negatives (All)					
1. Enterobacteriaceae					
2. Pseudomonas spp.					
3. Other					
II. OBLIGATE ANAEROBES (TOTAL)					
A. Bacteroides spp.					
B. Clostridium spp.					
C. Other					
III. FUNGI (TOTAL)					
A. Candida spp.					
B. Other					
IV. VIRUSES (TOTAL)					
V. Other (TOTAL)					
Polymicrobial cases as percent of all cases					
Total number of organisms/total number of poly-microbial cases					

**S. epidermis, Micrococci spp. & others; excludes coagulase-positive staphylococci

Data Definitions for Table 2

The basic information for tables 2a and 2b are the same. Table 2a displays data for patients meeting the mono-microbial definition and table 2b displays data for patients meeting the poly-microbial definition.

1. General information
   a. Percents will be displayed to 1 decimal
   b. Unless otherwise specified, percentages are based on the total number of patients in the group column
2. Mono-Microbial (Table 2a)
   a. Data for this table will only be based on patients with a positive culture for the SSTBJ group with a single organism. The exception to this is for cases with ONLY both MRSA and MSSA or ONLY both VRE and StrepD; under either condition the case would be considered Mono-microbial. Note too that where there is a separate line for MRSA and MSSA the cases that have both and only these organisms (in the final assigned site/culture as selected below) are shown on the MRSA line.
   b. If there are multiple qualifying cultures of the same site for the SSTBJ group, use the last culture by date (during the admission period) to determine single versus multiple organisms.
   c. If there are multiple qualifying cultures of difference sites for the SSTBJ group, use the last culture by date (during the admission period) for the deepest site to determine single versus multiple organisms. The order of the deepest sites are:

      Bone
      Joint
      Device/Prosthesis
      Skin
      Extremity
3. Poly-Microbial (Table 2b)
   a. Data for this table will only be based on patients with a positive culture for the SSTBJ group with more than one organism. Note the exception to this in 2a above.
   b. If there are multiple qualifying cultures of the same site for the SSTBJ group, use the last culture by date (during the admission period) to determine single versus multiple organisms.
   c. If there are multiple qualifying cultures of difference sites for the SSTBJ group, use the last culture by date (during the admission period) for the deepest site to determine single versus multiple organisms. The order of the deepest sites are:

      Bone
      Joint
      Device/Prosthesis
      Skin
      Extremity
   d. Polymicrobial admission—number of cases with poly-microbial culture. Percent is determined for the total number of poly-microbial cases divided by the total number of cases with a positive culture.

Appendix A: SSTBJ Diagnosis Codes: Sample List

CODE	DESCRIPTION	GROUP
99832	Disrupt External Op Wound	Surgical Wound (2a)
99851	Inf Postoperative Seroma	Surgical Wound (2a)
99859	Other Postop Infection	Surgical Wound (2a)
99883	NonHealing Surgical Wound	Surgical Wound (2a)
9964	Malf Int Orthped Dev/Gr	Prosthesis/Device (2b)
99640	MechComp Int Orth Dev NOS	Prosthesis/Device (2b)
99641	Mech Loosen Prosth Joint	Prosthesis/Device (2b)

Appendix B: Cultures

CODE	DESCRIPTION	NOTES
4001	Bone Culture	Includes any bone culture
4002	Device/Prosthesis Culture	Includes drainage tubes as device/prosthesis
4003	Joint/Muscle Culture	
4004	Skin Culture	Includes wound cultures
4005	Extremity Culture	Nonspecific sources such as leg, arm without further information to classify elsewhere

Appendix C: Organism Groupings

TABLE 2 GROUP	GROUP DESCRIPTION	CODE	ORGANISM DESCRIPTION	GENERAL CATEGORY
I A 1	S. aureus	4777	Staph. aureus	Gram positive
I A 1	S. aureus	4778	Meth Resist Staph. aureus	Gram positive
I A 2	Coagulase Neg Staph	4779	Staph. epidermis	Gram positive
I A 3 a	Enterococcus	4782	Vanc Resist Enterococcus	Gram positive
I A 3 a	Enterococcus	4786	Strep D	Gram positive
I A 3 b	Other Streptococci	4781	Strep ex B	Gram positive
I A 3 b	Other Streptococci	4783	Strep B	Gram positive
I A 3 b	Other Streptococci	4784	Strep A	Gram positive
I A 3 b	Other Streptococci	4785	Strep C	Gram positive
I A 3 b	Other Streptococci	4787	Strep F	Gram positive
I A 3 b	Other Streptococci	4788	Strep G	Gram positive
I A 3 b	Other Streptococci	4789	Strep Non Group	Gram positive
I A 4	Other Gram Positive	4753	Listeria	Gram positive
I A 4	Other Gram Positive	4757	Corynebacterium	Gram positive
I A 4	Other Gram Positive	4775	Streptobacillus	Gram positive
I B 1	Enterobacteriaceae	4801	Enterobacter	Gram negative
I B 2	Pseudomonas	4773	Pseudomonas	Gram negative
I B 3	Other Gram Negative	4751	Acinetobacter	Gram negative
I B 3	Other Gram Negative	4755	Providencia	Gram negative

Appendix D: Data Sets and Programs

SAS data sets created specifically for this project are located on <insert server name> in <insert directory information>. SSTBJ Data Mart is located on <insert server name>.

Data Sample List

DATA SET NAME	KEY/CONTENTS	CREATION/UPDATE PROGRAMS/SOURCES
Pat0203	Basic patient data for all eligible SSTBJ cases	Patlistsetup.sas TablemodsMMDDYY.sas
Cult0203	Culture/Organism data for eligible SSTBJ cases	Extract done by S. Smith from master culture data set
DxPat0203V DxPat0203tab	Normalized view/table of all diagnosis codes by case	Dxpat0203createview.sas (uses table Dxseq30)
Hosplist	Hospital/SSTBJ discharge date range	Hospitals and date ranges selected via iterative review of spreadsheets): Hospmosstbjdx_02.xls Hospmosstbjdx_03.xls Hospmolabvs.xls Hosptlistinitsetupvw.sas

APPENDIX D: Meaningful Use

The Medicare and Medicaid EHR Incentive Programs provide financial incentives for the "meaningful use" of certified EHR technology to improve patient care. To receive an EHR incentive payment, providers have to show that they are "meaningfully using" their EHRs by meeting thresholds for a number of objectives. CMS has established the objectives for "meaningful use" that eligible professionals, eligible hospitals, and critical access hospitals (CAHs) must meet in order to receive an incentive payment.

The Medicare and Medicaid EHR Incentive Programs are staged in three steps with increasing requirements for participation. All providers begin participating by meeting the Stage 1 requirements for a 90-day period in their first year of meaningful use and a full year in their second year of meaningful use. After meeting the Stage 1 requirements, providers will then have to meet Stage 2 requirements for two full years. Eligible professionals participate in the program on the calendar years, while eligible hospitals and CAHs participate according to the federal fiscal year.

REQUIREMENTS FOR STAGE 1 OF MEANINGFUL USE

Meaningful use includes both a core set and a menu set of objectives that are specific to eligible professionals or eligible hospitals and CAHs. For eligible professionals, there are a total of 24 meaningful use objectives. To qualify for an incentive payment, 19 of these 24 objectives must be met:

- 14 required core objectives
- Five objectives chosen from a list of 10 menu set objectives

For eligible hospitals and CAHs, there are a total of 23 meaningful use objectives. To qualify for an incentive payment, 18 of these 23 objectives must be met:

- 13 required core objectives
- Five objectives chosen from a list of 10 menu set objectives

CMS provides Meaningful Use specification sheets that bring together critical information on each objective to help you understand what you need to do to meet the program requirements. Each specification sheet covers a single eligible professional core or menu set objective in detail, including information on:

- Meeting the measure for each objective
- How to calculate the numerator and denominator for each objective
- How to qualify for an exclusion to an objective
- In-depth definitions of terms that clarify objective requirements
- Requirements for attesting to each measure

REQUIREMENTS FOR STAGE 2 OF MEANINGFUL USE

On September 4, 2012, CMS published a final rule that specifies the Stage 2 criteria that eligible professionals (EPs), eligible hospitals, and critical access hospitals (CAHs) must meet in order to continue to participate in the Medicare and Medicaid Electronic Health Record (EHR) Incentive Programs. All providers must achieve meaningful use under the Stage 1 criteria before moving to Stage 2.

CORE AND MENU OBJECTIVES

Stage 2 uses a core and menu structure for objectives that providers must achieve in order to demonstrate meaningful use. Core objectives are objectives that all providers must meet. There is also a predetermined number of menu objectives that providers must select from a list and meet in order to demonstrate meaningful use.

To demonstrate meaningful use under Stage 2 criteria:

- EPs must meet 17 core objectives and three menu objectives that they select from a total list of six, or a total of 20 core objectives
- Eligible hospitals and CAHs must meet 16 core objectives and three menu objectives that they select from a total list of six, or a total of 19 core objectives

APPENDIX E: CLINICAL QUALITY MEASURES

REPORTING ON CLINICAL QUALITY MEASURES

In addition to meeting the core and menu objectives, eligible professionals, eligible hospitals, and CAHs are also required to report clinical quality measures.

- Eligible professionals must report on six total clinical quality measures: three required core measures (or three alternate core measures) and three additional measures (selected from a set of 38 clinical quality measures).
- Eligible hospitals and CAHs must report on all 15 of their clinical quality measures.

CLINICAL QUALITY MEASURES FOR 2014 AND BEYOND

All providers **are required to report on CQMs** in order to demonstrate meaningful use. Beginning in 2014, all providers, regardless of their stage of meaningful use, will report on CQMs in the same way.

- **EPs must report on nine out of 64 total CQMs**
- **Eligible hospitals and CAHs must report on 16 out of 29 total CQMs**

In addition, all providers must select CQMs from at least three of the six key healthcare policy domains recommended by the Department of Health and Human Services' National Quality Strategy:

- Patient and family engagement
- Patient safety
- Care coordination
- Population and public health
- Efficient use of healthcare resources
- Clinical processes/effectiveness