Institute for Quality Improvement
5250 Old Orchard Rd, Suite 250
Skokie, IL. 60077

IQI Insights Volume 2, Number 2, Spring 2010

Cleaning and Analyzing Data

A Note to the Reader:

IQI Insights is a series of brief informational pieces from the AAAHC Institute for Quality Improvement. Our focus is on enhancing quality and safety through educational activities. In this series, we hope to provide you with the opportunity to learn more about basic issues and concepts associated with quality improvement in ambulatory health care. These short documents are not meant to provide in depth or complete information; however, we hope that they will increase your comfort with these topics and perhaps, lead you to seek additional information. We welcome your feedback.

Sincerely,

Naomi Kuznets, PhD, Managing Director
AAAHC Institute for Quality Improvement
nkuznets@aaahc.org

Introduction

This *IQI Insights* will focus on cleaning and analyzing data (please see AAAHC Standard 5.II.B. 5). This edition of *Insights* is designed to make these concepts more accessible to the research or statistically "challenged" reader; it is not exhaustive.

Where Do Cleaning and Analyzing Data Come into the Quality Improvement Process?

In the Quality Improvement (QI) process, you are asked to identify a potential important problem or issue (AAAHC Standard 5.II.B.1), set a performance goal (5.II.B.2), decide how to measure (develop questions to gather information that will answer whether you have a problem or not, how big a problem you have and possible causes of the problem) (5.II.B.3), and use the measures to actually collect information (data) (5.II.B.4). Now you need to "do something" with the data you have collected, but you don't know where to start. Please note: the discussions below will also apply if/when you have implemented a corrective action (AAAHC Standard 5.II.B. 7) to respond to a problem you have found on initial measurement and need to re-measure (5.II.B.8).

Cleaning Data

One of the first things you will need to do with your data is called "cleaning." This cleaning can be as tedious as the cleaning we do at home. How does one "clean" data?

Missing Values

1) You need to review your data to see whether there are any missing values (i.e. missing information or missing answers to your questions [measures]). For example, if you are using data collected from a survey and for one survey respondent, for one of the questions, no option is chosen, then you have missing data.
2) Where you have missing information, you must see whether there is an alternative data source that can help you accurately fill in the missing information. From the example above, you may be able to go back to the survey respondent and ask for an answer to the question with no option chosen. The information may also be available from another source, such as the respondent's medical chart.
3) Once you have filled in what missing information that you can, you need to decide whether you have enough information to go ahead.
    a. If the answer to a question is missing from one respondent's survey and there are 35 surveys collected, you are missing less than 3% (1/35) of the response to the question. When 5 respondents out of 35 don't answer the question, then you are missing slightly less than 15% (5/35) of the response to the question. This is important to at least note in your results; it may indicate several different issues, including respondents' fear of offending you, the lack of a choice of an answer that reflected their answer to the question, etc.
    b. What do you do with the survey (or other measurement tool) that is partially or almost totally incomplete? Here you must use your judgment to decide whether enough important data (data designed to answer your QI question[s]) is missing to warrant excluding this survey. For example, if the only information completed on a patient satisfaction survey is information on the patient's demographics (age, gender, race/ethnicity, etc.) and no information is completed with regard to the patient's experience or rating of care, the survey should most likely be noted in your report, but excluded from analyses.

Unlikely or "Outlier" Values

Another important aspect of "cleaning" data is checking for unlikely or "outlier" values in your data.

1) Contradictory information is one aspect of this issue. Examples include: pregnant males or patients who have rated all specific aspects of their care in the most positive manner (a "5" on a scale of 1-5, with 1= worst and 5 = best) but give the most negative overall rating of their care ("1" on the same scale). These answers can indicate poor respondent attention to the survey question, missing important issues in your measurement tool, or a confusing measurement tool.

2) Another aspect of this issue is the "outrageous" or "outlier" value. A statistical definition of an "outlier" is a value that is more than two standard deviations [1] from the mean (average—see bullet 3 toward the bottom of the page, for information on calculation of this value); in other words, the value you see is "way out there" or "outrageous." Examples include: patients who are less than 4 feet in height or administration of a probably lethal dose of a medication. These can be the result of respondent carelessness or inability of respondents to accurately calculate or convert information, or these may be real values—you need to check.

Analyzing Data

Once you have "cleaned" your data, you are ready to analyze or try to understand what your data mean. At this point, you need to remember one of the most important differences between "quality improvement" activities and "research," such as what you would read in peer-reviewed journals (*JAMA*, etc.). The data you are collecting for your QI activity are most likely not of the quality (accuracy/completeness) or quantity (sample size/ability to accurately represent your population) as that of peer-reviewed journal research. This is actually a good thing for those of you who are not comfortable or enthusiastic about using complex statistical analyses, because your data most probably will not warrant such analyses. For those of you who are disappointed by this, you should remember that most statistical analyses depend on a sufficiently large and representative sample, as well as a heterogeneous population, whose attributes fall on a normal, bell shape curve—you are not likely to see this in your QI data.

Sample Data

Patient Number	Height (inches)	Weight (pounds)	Asthma Follow Up Instructions (Yes or No)	Pre-Procedure Time (Minutes)
1	65	170	Yes	29
2	70	155	Yes	23
3	63	125	Yes	34
4	72	160	Yes	24
5	60	105	Yes	49
6	69	160	Yes	66
7	70	145	Yes	18
8	64	175	No	42
9	77	200	Yes	28
10	64	165	Yes	19
Range	60-77	105-200		18-66
Median	67	160		28.5
Mean (Average)	67.4	156		33.2

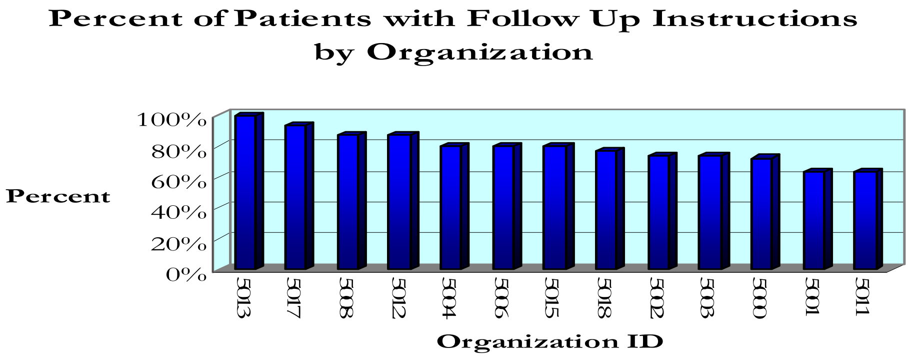
This leaves us with analyses most of us learned in grade or high school:

1) *Ranges*: this is the lowest and the highest value for a measure. Using the "Sample Data," if you put the values in order, the range for *Pre-Procedure Time*, it is 18 to 66 minutes.
2) *Median*: when you arrange your values for a certain measure from lowest to highest, this is the number in the middle. If you have an even number of values and the two in the middle are not the same, then you need to find the middle number between them. Using the "Sample Data" for *Height*, the middle numbers are 65 and 69 inches, so 67 inches is the median.
3) *Mean or Average*: if you add up all the values for a measure and divide by the number of values, you get the mean or average. Using the "Sample Data" for *Weight*, the sum of the values is 1560 pounds and this is divided by 10 (the number of patients for whom you have collected this information), so the mean or average is 156 pounds per patient.
4) *Percent*: when you express a proportion or decimal as a part of 100, this is a percent. For example, 9 of the 10 (or 0.9) patients measured for receipt of *Asthma Follow Up Instructions*, did receive instructions. Nine out of 10 or 0.9 expressed as a part of 100 is 90 out of 100 or 90%.

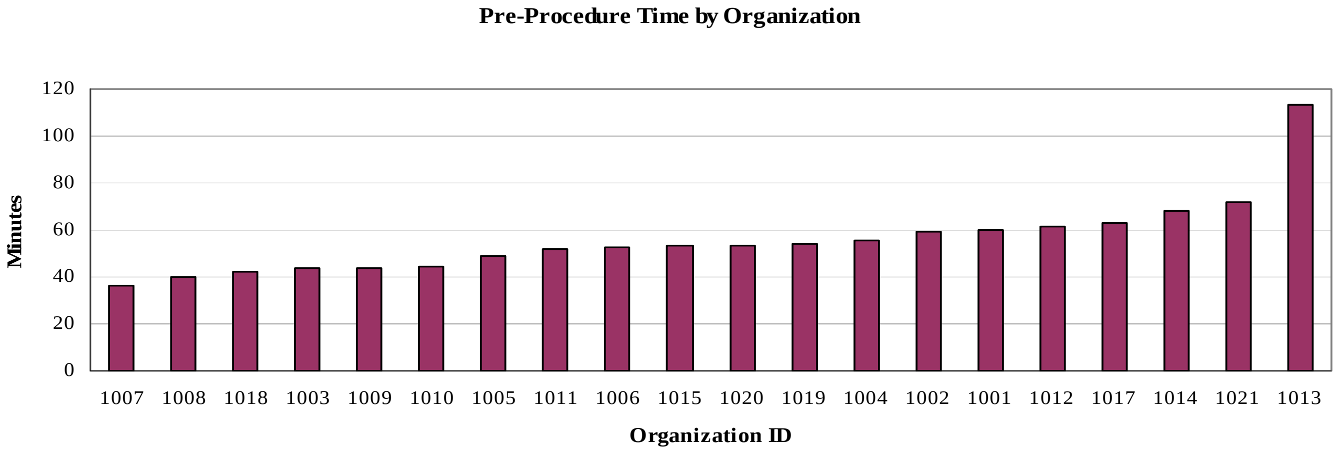
Please note that you can use Microsoft Excel to calculate medians [MEDIAN], averages [AVERAGE], and other statistical analyses (such as calculating the standard deviation [STDEVP] or correlation coefficient [CORREL]) using the Insert Function (*fx*) command to insert functions (formulas with information about which values to include). [2, 3]

What Do Your Analyses Tell You?

The benchmark graph below is from the 2008-2009 AAAHC Institute *Asthma Management in College Health* study. The National Heart Lung and Blood Institute asthma guidelines recommend regular follow-up visits at 1-6 month intervals in order to maintain control and consider appropriate step-down therapy. The overall combined (aggregate) median was 80%. The graph shows that if our Sample Data organization from the last page kept up their good work (as did Organization 5013, 5017, 5008, and 5012), they should probably *not* devote a lot of additional resources trying to comply with the NHLBI guideline. For organizations at the other end (right) of the graph, this appears to be an issue worth pursuing.



Below find benchmark graph from the 2004 AAAHC Institute *Myringotomy with Tympanostomy Tube Insertion* study. Some consultants may suggest that all patients should be seen within 30 minutes of arrival. The overall combined (aggregate) median and average overall pre-procedure time for those below is 53 minutes. The graph shows that our Sample Data organization from the last page, if they kept up their good work (as did organizations 1007, 1008, 1018, 1003, 1009, 1010, and 1005, below), they shouldn't spend a lot of additional resources on this (despite consultants' suggestions). For organizations at the other end (right) of the graph, this appears to be an issue worth pursuing.



Additional References—*please note: references to web sites or products are not endorsements.*

[1] For an example of how to calculate a standard deviation, please see: http://hubpages.com/hub/stddev

[2] For more information on using statistical functions in Microsoft Excel, please see: http://office.microsoft.com/en-us/excel/HP052030661033.aspx

[3] For an example of how to calculate a correlation coefficient, please see: http://www.bized.co.uk/timeweb/crunching/crunch_relate_illus.htm