EMERGENCY PHYSICIANS MONTHLY

October 2010 | *Volume 17, Number 10* | www.epmonthly.com

NIGHTSHIFT

Night Caller

by Mark Plaster, MD

When the phone rang at 4:30 in the morning, instead of sleepily reaching for the bedside phone, my wife literally jumped out of the bed. With an elderly parent, she is always worried that such late night calls mean bad news. I, on the other hand, have a more sanguine view of life, and try to avoid any such thoughts that might interrupt a good night's sleep. I had just fin-

continued on page 19

IN THIS ISSUE

Taking Obamacare to Court

What will come of the recent constitutional challenge to the Patient Protection and Affordable Care Act? **by Kevin Klauer, DO & Mark Plaster, MD, JD**

Even before President Obama's ink was dry from signing the Patient Protection and Affordable Care Act (PPACA) into law, opponents gave notice that they would challenge the constitutionality of the law in the courts. The most recent challenge, on July 26th, a suit was filed by a private citizen, Matt Sissel. The defendants in this suit are two well known individuals: Kathleen Sebelius, Director, Health and Human Services and Timothy Geithner, United States

continued on page 18

Is Press Ganey Reliable?

Part II: Small Samples Create Questionable Results

by William Sullivan, DO, JD & Joe DeLucia, DO

How many patient satisfaction surveys are necessary to obtain a statistically reliable look at the performance of hospitals and health care providers? Press Ganey states that only 30 survey responses are needed to draw meaningful conclusions, although they prefer to have at least 50 responses before analyzing the data. We asked Dr. Eric Armbrecht, a statistician and Assistant Professor for St. Louis University's Center for Outcomes Research and Dana Oliver, a biostatistician at St. Louis University if they agreed.

Dr. Armbrecht suggested that analyzing only 30-50 responses would lead to unacceptably wide confidence intervals and would substantially limit the generalizability and use of the data obtained, regardless of whether 3,000 or 10,000 patients were surveyed. Dr. Armbrecht explained that low response rates could create confidence intervals as wide as 50%, which could be similar to just flipping a coin to determine whether the data is representative of an entire population's perceptions. Breaking down those same 30-50 responses in an attempt to analyze satisfaction scores of individual physicians would create even less reliable results as the number of responses per physician would be even less. Ms. Oliver also disagreed with Press Ganey's assertion that 30 or 50 responses would result in statistically sound data, noting that those numbers could be "arbitrarily chosen" by some survey methdologists.

continued on page 20

INSIDE
Why statistically, patient sat surveys shouldn't be used to compare different EDs

THE SUPER BOWL DOC

For more than 20 years Ricardo Martinez has coordinated medical care at America's most popular sporting event.

OH HENRY

"Last Scene of All"

If physicians can't give sound end-of-life advice to patients and families, who will?
by Greg Henry, MD

The principle issue in this country today, with regard to medicine, is not any particular form of research. It is not any particular operation. It is what are we going to supply to elderly people where there are limited resources and a shrinking base of people to pay the bills. This is more than symbolic. It is a day-to-day

continued on page 39

THE LITERATURE

When Tests Cry Wolf

A test might be "sensitive," but is it "specific" enough to be valid and useful?
by Jerome Hoffman, MD

Many studies try to evaluate the use of tests to find disease of one sort or another. "Tests" are of course not limited to laboratory studies or X-rays, but for the purposes of this type of research may include a given historical finding, an abnormality on physical examination, a set of "high-yield criteria" or anything we

continued on page 21

VAMPIRES IN THE ED? THE SCIENCE BEHIND PORPHYRIA

Are Press Ganey Statistics Reliable?

◀ *from cover*

How many responses are necessary in order to have statistically reliable data? The answer depends upon the size of the sample population. Assuming a margin of error of 4% (which is double the margin of error that Press Ganey would like to use) and assuming a statistical standard 95% confidence interval, the minimum sample sizes that Dr. Armbrecht recommended for populations of 1000, 2500, 5000, and 7500 would be 375, 484, 536, and 556 respectively. He noted how the response rate tends to flatten out with larger sample sizes and cautioned that these response rates would only apply to "yes/no" questions (such as whether or not a doctor was "very good"). In order to measure the validity of rating scales (such as those from 1-5), the calculations become somewhat more difficult and are dependent upon the standard deviation in the sample population. Dr. Armbrecht gave an example that using a 1-5 scale with a standard deviation of 0.7 and a margin of error of 10% (which is five times higher than Press Ganey seeks), 188 responses would be needed in order to reliably estimate the responses from the general population. Dr. Armbrecht recommended online statistical calculators such as those available at Creative Research Systems (http://www.surveysystem.com/sscalc.htm) to help determine the statistical significance of most data.

Aside from low response rates, Dr. Armbrecht and Ms. Oliver described additional problems that can occur when using a 1-5 scale in satisfaction surveys.

If hospital administrators seek to be at or above the 90th percentile in satisfaction scores, asking patients to grade performance on a 1-5 scale essentially creates a system with one passing grade and four failing grades. If patients are not aware that a score of "4" is a failing grade, the data that they provide may be misinterpreted when being analyzed. In addition, patients may perceive a small relative difference between a grade of "4" and "5" on a survey, but may perceive a larger relative difference between a "3" and a "4" on the same survey, creating a system in which they grade "so-so" care with the same score as "just less than perfect" care. Finally, with small sample sizes, one unhappy customer can turn many "passing" grades into failing grades. Four patient scores of "perfect" fives can be brought down to "failing" fours by one extremely unhappy patient who grades a provider or hospital with scores of all zero. Our experts noted that a simple way to avoid these analytical problems was to create a dichotomous scoring system with "yes-no" questions. For example, "Did your care meet your expectations?"

Glossary of Statistical Terms

MEAN: The average of all the responses.

MEDIAN: The middle value in all the responses when those responses are arranged in numerical order. The closer that the mean and the median get to each other, the less variance there is in the data.

DICHOTOMOUS DATA: Contains only two possible choices, such as whether the light was on or off.

NON-DICHOTOMOUS DATA: Consists of multiple possible values, such as rating scales used in satisfaction surveys.

NORMAL OR GAUSSIAN DISTRIBUTION: Another way of describing a typical bell curve.

STANDARD DEVIATION: The square root of the variance in a data set. Low standard deviations mean that the data points are close to the mean while high standard deviation values mean that the data points are spread out over a large range of values. When there is a normal distribution of data, about 68% of the data values will fall within one standard deviation of the mean and about 95% of the data values will fall within two standard deviations from the mean.

CONFIDENCE INTERVAL: A measure of survey reliability. The narrower the confidence interval, the more reliable the survey results. A confidence interval of 95% is the conventional standard in medical and social science research and reflects a high likelihood that the sample data reflects the population from which it was sampled.

Clarifying Terms

Press Ganey's literature contains several other statistical terms that our experts felt it was important to understand when analyzing the utility of patient satisfaction scores.

The "standard error of the mean" is the standard deviation of a sample population's mean. Ms. Oliver noted that before performing any type of statistical testing, it is a good idea to first plot a histogram of multiple sample responses to determine whether survey data will be distributed in a normal bell curve pattern. If the survey responses are not distributed in a bell curve pattern, conclusions cannot be drawn from the data – unless the variability of the data is low.

Press Ganey literature relies on the "central limit theorem" in justifying a reliance upon sample sizes as low as thirty. Ms. Oliver explained that the central limit theorem holds that the mean and median scores from very large survey samples tend to form a typical bell curve. In most cases, the central limit theorem only applies if there is a similar distribution of variables in each survey. Because patient satisfaction survey samples from specific hospitals are generally not large and because the surveys do not always have a similar distribution of variables, the central limit theorem probably would not apply to satisfaction survey data.

Analysis of survey results depends in part on the "margin of error" of the survey data. Margin of error is used to express the confidence with which survey responses can be relied upon when an entire survey population is incompletely sampled. For example, suppose that five percent of a sample population is surveyed and one question has a mean score of 50. If the margin of error for the question is 30, then the actual value for the response in the sample population could be anywhere between 20 and 80 (the mean score of 50 plus or minus 30). Dr. Armbrecht stated that a good estimate of a margin of error is given by the formula 1/[square root of the number of participants in the sample size] (Niles, 2006). In other words, for a sample size of 100, the margin of error would be roughly 10% and for a sample size of 9, the margin of error would be roughly 33%. Achieving Press Ganey's goal margin of error of 2% or less would require a sample size of approximately 2500.

Understanding Survey Limitations

So are satisfaction surveys a useful tool for assessing the quality of medical care? Dr. Armbrecht compared analysis of survey data to sampling a pot of soup.

If you want to see how good the soup in a pot tastes, first the ingredients in the pot must be well mixed. The "mixing" of the soup is analagous to obtaining completely random data from a sample population. If you only mix the top layers of the pot, you might not get the beans and pasta on the bottom of the pot, so your sample taste will not be representative of the true flavor of the soup. Similarly, failure to completely randomize data samples by excluding certain segments in a population (such as admitted, transferred or LWOBS patients) significantly increases the likelihood that the results will be inaccurate.

If the soup is fully mixed, but you only taste a drop or two of soup, you probably won't get a good flavor for the soup, either. Similarly, small sample sizes from a large population are likely to provide misleading data.

Once an appropriate sample is taken, surveys can only be used to determine whether there has been a change in the sample population. Using the soup analogy, you tweak the recipe by adding or changing ingredients and take another sample to see if people like the new recipe better. Surveys can only be used to measure how the soup in a single pot is changing over time.

Sometimes survey data can be misused, though. For example, sampling the soup in two different pots can't tell you whether one soup is better than another soup or whether one ingredient is better than the same ingredient in a different pot. Satisfaction survey statistics likewise should not be used to compare and rank different hospitals or different health care providers. Dr. Armbrecht noted that a 90% ranking at one hospital cannot be deemed better or worse than a 70% ranking at a different hospital. The demographics and variance in patient populations being sampled don't allow such a comparison as it is more likely that variables independent of the services provided (such as patient literacy, lack of forwarding address, language barriers, payment issues, and population homogeneity) will have an effect on the data being sampled. In other words, taking the staff from a hospital with 90% satisfaction score and placing them into a different hospital would probably not create a 90% satisfaction score in the new hospital. The only information that satisfaction surveys can provide is a determination whether a specific hospital or a specific provider at a hospital is getting better or worse over time. In order for even that determination to be made, the sample sizes must be large enough to be statistically significant.

What are the takeaway points about analysis of satisfaction survey data?

First, small sample sizes can lead to significantly unreliable data. Last month, we showed how small sample sizes resulted in a 99% change in a hospital's percentile rank in just two months. Simply put, small response sizes lead to inaccurate results.

Second, when sample sizes are large enough, satisfaction surveys can be an important tool to gauge and improve patients' perception of the medical care they receive. However, using survey data to compare one hospital to another or to compare one provider to another is a misuse of survey data and is likely to create misleading and unreliable results.

For a full list of references, go to www.epmonthly.com