Behavior and Analysis of 36-Item Short-Form Health Survey Data for Surgical Quality-of-Life Research

Vic Velanovich, MD

Hypothesis: Data from the 36-Item Short-Form Health Survey (SF-36) do not follow a normal distribution and should not be analyzed using parametric techniques. A novel type of analysis, top-box analysis, may add to the interpretation of these data.

Design: Review of SF-36 data from preoperative and postoperative patients.

Setting: Tertiary care hospital and clinic.

Patients: One thousand randomly selected preoperative and postoperative patients with a variety of surgical diseases completed the SF-36 (8 domains: physical functioning, role physical, role emotional, bodily pain, vitality, mental health, social functioning, and general health). The best possible score was 100; the worst possible score, 0. One item assessed “health transition.” The best score was 1; the worst score, 5. The health transition item and each domain were analyzed for mean with standard deviation, median, mode skewness, kurtosis, and normality. A “top-box” assessment was done by determining the frequency of patients scoring 100 in each domain or 1 in the health transition item. In addition, preoperative and postoperative scores were compared.

Results: The results for all 1000 questionnaires demonstrated that none of the domains had data that followed a normal distribution. The means, medians, and modes were different. Five domains had the mode and median at the top box.

Conclusions: The SF-36 data did not follow a normal distribution in any of the domains. Data were always skewed to the left, with means, medians, and modes different. These data need to be statistically analyzed using nonparametric techniques. Of the 8 domains, 5 had a significant frequency of top-box scores, which also were the domains in which the mode was at 100, implying that change in top-box score may be an informative method of presenting change in SF-36 data.

Arch Surg. 2007;142:473-478

Quality-of-life (QoL) measurement has become an increasingly used endpoint in assessing surgical outcomes. Many surgical studies have used QoL instruments to assess changes in QoL. A MEDLINE search of 12 leading US surgical journals, using the keywords quality of life, from 1996 to October 2006 yielded 466 articles. Although there are a variety of QoL instruments available, one of the most commonly used is the 36-Item Short-Form Health Survey (SF-36). Once again, a MEDLINE search of these journals using the keyword SF-36 yielded 92 articles. When looking at the keyword SF-36, more than 4000 articles are found. The SF-36 is a generic QoL instrument that measures 8 domains of QoL; these 8 domains can be reduced to 2 summary scores—a physical component and a mental health component. Because of its popularity, the SF-36 is considered one of the leading QoL instruments worldwide, and has been translated into dozens of languages.

A review of surgical QoL studies has found that there were several deficiencies in the conduct of these studies. One of the most common problems was inappropriate statistical analysis. The proper statistical analysis of data is essential in interpreting the results of any study. Commonly, data from the SF-36 have been presented as means with standard deviations or standard errors of the mean. The basic assumption of these studies is that the data follow a normal (gaussian) distribution, having a “bell-shaped” curve. However, many of these studies did not perform the statistical tests needed to determine if, indeed, the data follow the normal distribution necessary to use this type of statistical analysis.

The purpose of this study was to analyze the behavior of SF-36 data in a group of preoperative and postoperative surgical patients to determine the most appropriate statistical methods of analysis.

Author Affiliation: Division of General Surgery, Henry Ford Hospital, Detroit, Mich.
THE SF-36 GENERIC QoL INSTRUMENT

The SF-36 measures 8 domains of QoL: physical functioning, limitations to physical activities because of health, such as self-care, walking, and climbing stairs; role physical, interference with work or daily activities because of physical health; role emotional (RE), limitations to work or daily activities because of emotional health; bodily pain, pain intensity and how this affects work in and out of the home; vitality, how full of energy the patient feels; mental health, overall emotional and psychological status; social functioning, how much health interferes with social interactions; and general health, overall evaluation of health. The scores are standardized so that the worst possible score is 0 and the best possible score is 100. In addition, question 2 addresses “health transition”: “Compared to one year ago, how would you rate your health in general now?” Possible answers are as follows: 1, much better now than 1 year ago; 2, somewhat better now than 1 year ago; 3, about the same as 1 year ago; 4, somewhat worse than 1 year ago; and 5, much worse than 1 year ago.

PATIENTS AND INSTRUMENT ADMINISTRATION

It is part of my practice to administer the SF-36, and other QoL instruments, depending on disease process, to all patients seen in consultation in the outpatient setting of my general surgical practice. The patients are given the questionnaire and are merely instructed to complete the questionnaire before the clinical encounter. No one helps the patient complete the questionnaire to avoid any physician-related bias in the answers. Patients may or may not undergo a surgical procedure. Generally, postoperative patients complete the questionnaire 6 to 12 weeks postoperatively, during a follow-up visit. It is my practice to provide long-term follow-up of patients with cancer and some patients with nonmalignant disease who require long-term follow-up. In these patients, the questionnaire is administered yearly. I score the instrument, recording the scores of each domain, without determining the physical or mental health component summary scores.

SELECTION OF QUESTIONNAIRES INCLUDED IN THE STUDY

A random sample of 1000 answered questionnaires was selected. Randomization was based on selection of the last digit of the patient’s medical record number from a random number table. This process was repeated until 1000 questionnaires were selected. One thousand was selected as an adequate sample size to leave no doubt as to the presence of skewness and kurtosis. Because of the random nature of selection, the sample included preoperative and postoperative questionnaires, and questionnaires from the same patient at different times could have been included. Because each questionnaire is linked to a patient, the patient’s diagnosis is known and was recorded.

STATISTICAL ANALYSIS

All statistical analysis was done using a statistical computer program. Initially, each questionnaire was reviewed for missing items (ie, questions not answered). Questionnaires with missing data were still included in the analysis, with domains that could be scored analyzed. The denominators changed according to how many domains had scores that could be analyzed. Data from each domain of the SF-36 were analyzed for the descriptive statistics of mean with standard deviation, median with complete range and interquartile range, mode, skewness, and kurtosis. The data were analyzed to determine fit to a normal distribution using the Wilk-Shapiro test. \( P < .05 \) was considered statistically significant for the data not fitting the normal distribution. Last, because the highest score achievable on the SF-36 is 100 for each domain and 1 for the health transition item, this score was considered the “top box.” The frequency of this top-box score was determined for each domain. The top-box score was inspired by its use in patient satisfaction measurements from Press Ganey Associates, Inc, South Bend, Ind (http://www.pressganey.com). Press Ganey Associates, Inc, provides a survey to measure patient satisfaction. These surveys are sent to the patient and returned to the company. The company reports include a summary score plus the distribution of the frequency of each answer selected—from poor to very good. The “very good” selection is the top box by this method. Press Ganey Associates, Inc, then compares the frequency of this top-box score with other institutions using the survey to determine the percentile rank of the institution or individual practitioner. It is hoped that this information will help identify areas in need of improvement.

RESULTS

Of the 1000 questionnaires evaluated, 541 were preoperative and 459 were postoperative. The percentage of missing items ranges from a high of 5.7% (in the social functioning and general health domains) to a low of 2.1% (in the physical functioning domain). Table 1 shows the summary statistics for each domain for the entire sample of 1000 questions. Using the Wilk-Shapiro test, none of the domains, or answer distribution for question 2, follow a normal distribution. The results of the Wilk-Shapiro test are statistically significant, proving a

<table>
<thead>
<tr>
<th>Domain</th>
<th>Mean (SD)</th>
<th>Median</th>
<th>Mode</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Top-Box Frequency, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Physical functioning</td>
<td>70.3 (28.0)</td>
<td>80.0</td>
<td>100</td>
<td>-0.75</td>
<td>2.42</td>
<td>19.1</td>
</tr>
<tr>
<td>Role physical</td>
<td>58.7 (43.8)</td>
<td>75.0</td>
<td>100</td>
<td>-0.35</td>
<td>1.35</td>
<td>46.2</td>
</tr>
<tr>
<td>Role emotional</td>
<td>74.5 (39.1)</td>
<td>100.0</td>
<td>100</td>
<td>-1.08</td>
<td>2.43</td>
<td>66.8</td>
</tr>
<tr>
<td>Bodily pain</td>
<td>62.4 (28.0)</td>
<td>62.0</td>
<td>100</td>
<td>-0.26</td>
<td>2.08</td>
<td>20.2</td>
</tr>
<tr>
<td>Vitality</td>
<td>53.6 (23.3)</td>
<td>55.0</td>
<td>50</td>
<td>-0.28</td>
<td>2.45</td>
<td>39.9</td>
</tr>
<tr>
<td>Mental health</td>
<td>73.9 (19.4)</td>
<td>80.0</td>
<td>92</td>
<td>-0.98</td>
<td>3.64</td>
<td>5.1</td>
</tr>
<tr>
<td>Social functioning</td>
<td>73.9 (28.2)</td>
<td>87.5</td>
<td>100</td>
<td>-0.84</td>
<td>2.68</td>
<td>39.9</td>
</tr>
<tr>
<td>General health</td>
<td>64.3 (22.1)</td>
<td>67.0</td>
<td>77</td>
<td>-0.48</td>
<td>2.55</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Abbreviation: SF-36, 36-Item Short-Form Health Survey.

*The Wilk-Shapiro \( P \) value was <.001 for all domains.
nonnormal distribution. The coefficients of skewness are all negative, implying a long “tail” toward smaller values than the mean. The coefficients of kurtosis are less than 3 (the value for a normal distribution) in 7 of 8 domains and question 2 and greater than 3 in 1 domain. Values less than 3 imply that the “shoulders” of the curve are broader than what would be expected from a normal distribution, while values greater than 3 imply a tall “peak” of the curve with narrow shoulders. The value of 3 is expected for data following a normal distribution.

**Figures 1, 2, 3, 4, and 5** show illustrative comparisons of preoperative with postoperative histograms of the number of each answer of question 2 (health transition) and the scores of physical functioning, role physical, RE, and vitality domains, respectively. Visually, none of the shown or not shown histograms follow a normal distribution, although the vitality domain seems the most like a bell-shaped curve. Similar descriptive statistics were done on each of these data sets, as with the overall data sets. The results also demonstrated different means, medians, and modes, negatively skewed data, with kurtosis different than 3, and significant Wilk-Shapiro tests, proving that none of the data sets follow a normal distribution.

Table 1 also presents the frequency of top-box scores. In 5 of the 8 domains, many patients scored 100 (the top box). In only 3 domains was the frequency of a top-box score less than 10%. Also, the domains with a higher frequency of top-box scores were also the domains in which the mode was at 100. However, when comparing preoperative with postoperative top-box scores, the frequencies are different (Figures 1-5). Once again, these differences are more apparent in those domains in which the top box was the mode.

**COMMENT**

This study demonstrated that, in a general surgical population, the behavior of SF-36 data does not follow a normal distribution. In fact, the role physical and RE domains have more of a U shape than a bell shape (Figure 3 and Figure 4, respectively). This finding is significant for surgical QoL research. Data that do not follow a normal distribution should not be analyzed with parametric statistical techniques. The presentation of nonnormal data as means with standard deviations or standard errors of the mean misrepresents the data; analysis of such data with parametric techniques, such as a t test, is inappro-
These types of data should be analyzed using non-parametric techniques. Ultimately, the main point of this study is that it is essential to assess that continuous data follow a normal distribution before using parametric statistical analysis.

The natural question to ask is the following: Are the results of this study unique to my patient population or to patients in general? The developers of the SF-36 have done similar studies in the general population. Their results are similar. Table 2 presents their data from a sample of 2474 patients.
individuals. Although the distribution of the curves is shifted to the right (ie, means, medians, and the frequency of top-box scores are higher), the general shape of the curves is the same. Specifically, there are means and medians greater than 50, with a long tail to the left (ie, lower scores). Therefore, rather than an aberration, the distribution of SF-36 scores in the surgical population is consistent with the general population, although somewhat lower, as would be expected in patients requiring an operation.

The authors of the SF-36 recognized limitations in the 0 to 100 scoring scale. Their specific concern was related to how the scores between domains would be compared. Therefore, they developed a “norm-based” scoring system in which the mean, by definition, would be 50 and the standard deviation would be 10. Using a formula, the individual patient’s score can be transformed into this norm-based score. The authors make clear that the purpose of norm-based scoring is to provide a basis for meaningful comparison of scores between each domain and to compare the score of a single patient with that of the general population. This can beg the following question: Because these scores are “normalized,” does that imply that data using these scores will follow a normal distribution? In the study by Ware and Kosinski, their Table 10.2 provides the norms for the general US population. By definition, the mean is 50 and the standard deviation is 10. However, in none of the 8 domains is the median 50, although the median is within 1 point in the general health domain. In all domains, the top of the range (ceiling) is closer to 50 than the bottom of the range (floor). This implies that the histograms of these normalized scores are skewed, with a long tail to the left, just as we see in the standard scores. Therefore, even if the researcher wants to use the population-based norms, the researcher cannot assume that the data will follow a normal distribution.

Other subtleties of the data also should be noted. Although presented as a scale from 0 to 100, in fact, not all numbers between 0 and 100 are available as scores. This is a matter of “precision.” The more values available, the easier it is to differentiate between levels of QoL. For example, consider the RE domain. There are only 4 possible scores (0, 33, 67, and 100). In fact, the mean of 74.5 is not available as a choice. Therefore, one can argue that the scores of the SF-36 are not continuous at all, but rather ordinal, much like cancer staging. We know, by definition, that stage 3 disease is worse than stage 2 disease, but is there a definition of stage 2.5 disease? Most would say that stage 2.5 disease in cancer has no meaning. By analogy, one can argue that scores in between the scores that can be actually obtained have no meaning. Noncontinuous (ie, ordinal) data should be analyzed using nonparametric techniques.

This study also demonstrates that there would be value in a top-box analysis. Top-box analysis is the separate statistical analysis of the highest possible scores achievable by a survey (ie, the best score). These data then become nominal. Top-box analysis would be of most value when the distribution of scores bunches toward the highest value. For example, in the health transition question (Figure 1) and the role physical (Figure 3), the RE (Figure 4), and social functioning domains, the difference in the frequency of the top-box score was more dramatic than the difference in the median scores. In fact, in the RE domain, the preoperative and postoperative medians are the same, yet it is the top-box frequency that is substantially different. It is probably not coincidental that those domains also had the fewest possible scores (ie, the least precision). Those domains for which the top-box score was most insightful were the domains in which the mode was also at the top box. As seen in the RE domain, if the mode and median are both at the top box, then the frequency of the top-box score may be the only way to determine if there is a change in QoL score.

In conclusion, the SF-36 is a valuable instrument in measuring QoL in a variety of patient populations, including surgical patients. Nevertheless, researchers and readers of these studies need to be aware of the nature of these data and need to use appropriate statistical techniques.

Accepted for Publication: January 5, 2007.

Correspondence: Vic Velanovich, MD, Division of General Surgery, Mailstop K-8, Henry Ford Hospital, 2799 W Grand Blvd, Detroit, MI 48202 (velаноv1@hfhs.org).

Financial Disclosure: None reported.

Previous Presentation: This paper was presented at the 114th Annual Scientific Session of the Western Surgical Association; November 14, 2006; Los Cabos, Mexico; and is published after peer review and revision. The discussions that follow this article are based on the originally submitted manuscript and not the revised manuscript.

REFERENCE


*(REPRINTED) ARCH SURG/VOL 142, MAY 2007 WWW.ARCHSURG.COM

©2007 American Medical Association. All rights reserved.
DISCUSSION

John A. Weigelt, MD, Milwaukee, Wis: Quality measures have been a topic of our meeting as they are now a part of our health care system today. The SF-36 is one of many tools used to measure health status. It has been shown multiple times to be a reliable and valid measurement. It is used to measure health status before and after an intervention such as a surgical procedure. It is also associated with a growing body of literature trying to explain its use and how to apply it to various investigations.

Currently, the simple fact is if we wish to assess quality outcomes, we will eventually be forced to use a tool such as the SF-36. As you do this, please think about this presentation. This paper must be consumed with its figures. It is a visual thing. And unless you visually digest the graphs, the manuscript will fall a little flat. Once the graphs are interpreted, the whole premise of the article becomes clear.

We have heard a statistical challenge to how we are analyzing our SF-36 output. Most studies use parametric tests which assume a normal distribution. Dr Velanovich clearly demonstrates that the data we collect with the SF-36 are not always normal in their distribution but have many nonparametric characteristics. It is this observation that will make this report memorable after today.

Now, whether this top-box approach is the correct approach will need to be assessed by health care researchers and their statisticians.

I have 3 questions. Since many report SF-36 data using parametric analyses, what made you look at the data this way? The second question, nonparametric options are many and each have their champions. What was your reasoning for selecting the top-box methodology? And finally, since the top box was not ideal for all the domains, are you using any other nonparametric analyses to assess other parts of the SF-36?

Dr Velanovich: I have read many articles and heard many talks with regard to the SF-36 and the data have frequently been presented as parametric in nature (ie, following a normal distribution with means and standard deviations). Every time I assessed my data, the distribution was never normal. I was becoming concerned that I was wrong. My goal was to definitively answer the question about normality, at least in my mind.

The second question was about nonparametric tests and the reason for using the top-box methodology. At our institution, we use Press Ganey scores for measuring patient satisfaction. One of the ways that the scores are presented is in this top-box fashion. As I was reviewing these reports, it occurred to me that many of the SF-36 scores are at the 100 level and, perhaps, this could be a method to assess change in the data. So, this was the inspiration of the top-box method. I also agree that I don't know whether long-term this is going to be helpful or not, but I believe it is worth exploring.

The last question was about the appropriateness of the top box for all domains. Dr Weigelt is correct. It is not ideal for all the domains. So I suspect it is just going to be 1 of many ways to assess by health care researchers and their statisticians.

Donald E. Low, MD, Seattle, Wash: The issue regarding application of the SF-36 for many of us relates to the specifics of individual patient populations. For some of us who deal with cancer patients, assessing quality of life before and after surgery is unlikely to be productive.

Patients often have just learned they have cancer before they get to our office. Asking them to assess quality of life at that moment and then comparing to the postoperative parameters is open to misinterpretation.

The real advantage of the SF-36 in many of our minds is the ability to assess postoperative parameters compared to general and sex-matched populations, to gauge whether these patients can undergo large operations and return to a quality of life comparative to the general population. Is this not 1 of the advantages of the SF-36 in its current form when assessing quality of life, especially in cancer patients?

Dr Velanovich: The summary scores attempt to standardize the scores to population controls. The reason why the Medical Outcomes Trust went that way was because of the problem they had with the distribution of the data. If you look at the original articles, they had the same or similar distribution of data that I show here. So I don't think there is a difference specifically with surgical patients compared to the general population. The summary statistics are used in order to try to avoid some of the statistical issues.

For a cross-sectional study, which is what you are recommending, summary scores are a very good way of presenting these types of data. The problem is in determining the change of scores for individual patients. I think this is more problematic. I agree with you, there are potential problems in a cancer patient, who clearly may have anxiety and other issues. Nevertheless, it may not reflect what their precancer baseline is, but it is a valuable pretreatment baseline.

Karen J. Brasel, MD, Milwaukee: As you have shown, there are so many nonparametric statistical tests available; one might work for 1 of the domains but another is going to be required for a separate domain. Since now norm-based scores for the SF-36 are available for all of the domains, might an easier approach be to change, rather than using the 0 to 100, use the norm-based scores and simplify the subsequent statistical analysis?

Dr Velanovich: I don't have an answer for you. I have not used the norm-based scores in my research. So I don't know how much of that will change. One of the issues, though, is this whole point about what is a minimally important difference? And the minimally important difference that the SF-36 looks at is generally changes that are between 5 to 10 points. And that is based on the original parameters.

Basil A. Pruitt, Jr, MD, San Antonio, Tex: I wonder if your random selection of 1000 tests didn't ensure skewness and kurtosis because you included test results representing heterogeneity of age, heterogeneity of disease, and both preoperative and postoperative assessments. So how do we interpret that? I think that if you had a specific disease, specific age group, and postoperative, and maybe even sequential, postoperative assessments, because now there are techniques to look at trajectory of recovery across time, that you would have reduced the variability that you are decaying.

Dr Velanovich: Ultimately, the point is that a researcher needs to test that his or her data follow a normal distribution prior to the use of parametric statistical tests.

And notice there actually is not that much difference between the shape of the curves between preoperative and postoperative. I haven't shown here, but when you separate out cancer patients from noncancer patients, although the means and medians and the curves shift, the shapes of the curves look pretty much the same.

Walter J. McCarthy, MD, Chicago, Ill: Because of the heterogeneity of the patients, you are going to have a wide spectrum of results. We reviewed a large number, more than 300, patients with intermittent claudication and found that, particularly with the physical function aspect of the SF-36, the distribution was quite normal. We were also able to compare pretreatment and posttreatment physical function scores in a meaningful way, showing an improvement. My point is, in a homogeneous group, the SF-36 data will be more normal.

Dr Velanovich: I applaud you for at least looking at it and taking that into consideration. Most of the time, though, even with homogeneous groups, such as with reflux patients, it still comes out this way.

Financial Disclosure: None reported.