ROBUSTNESS OF THE ACHIEVABLE BENCHMARK OF CARE METHOD

by

JEFF ARTHUR CAPOBIANCO

DISSERTATION

Submitted to the Graduate School

of Wayne State University,

Detroit, Michigan

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

2012

MAJOR: EDUCATIONAL EVALUATION & RESEARCH

Approved by:

_____

Advisor Date

_____

_____

_____

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

DEDICATION

To my whole family and especially my wife Laurel.

ACKNOWLEDGMENTS

I thank my committee members, especially Dr. Shlomo Sawilowsky and the late Dr. Donald Marcotte, for their advice and support. I also greatly appreciate the kindness, encouragement, and guidance provided by Dr. Mary Ruffolo, Dr. Marilyn Wedenoja, Dr. Kristen Barry, Dr. Fredrick Blow, and Dr. Laura Lein.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER 1: INTRODUCATION & BACKGROUND

The physician Ernest Codman (1869-1940) was a pioneer in the field of healthcare quality improvement and his advocacy for sharing information on healthcare provider performance has been said to have, "brought him mostly ridicule, poverty and censure" (Neuhauser, 2002, p.105). Currently, the measurement and reporting of healthcare provider performance occurs at both the individual clinician and aggregate provider levels and is monitored and compared by a variety of stakeholder groups including accrediting agencies, payers, policy makers, researchers, quality improvement teams, and patients. It took decades for Codman's advocacy to take hold and that occurred in large part because the United States Federal government created incentives to make the measurement and reporting possible. For example, in 2004 the U.S. Centers for Medicare & Medicaid Service (CMS) began to incentivize physicians and hospital systems to report on care provided and associated outcomes through increased reimbursements known as "pay for reporting"(Neuhauser, 2002, p. 104). Prior to 2004 this kind of information was only collected and reported to accrediting agencies and rarely made it into the public arena. Similarly, in 2005 the CMS began the public reporting of hospital process measures on the Hospital Compare website <http://www.hospitalcompare.hhs.gov> by 2015 the failure of a healthcare provider to report on care processes, and associated outcomes, will result in monetary penalties (Mather, Hettrich, & Nunley, 2011).

Although the idea of sharing quality information with stakeholders is no longer a topic that draws ridicule or censure studies on how best to measure, collect, compare, and report healthcare performance data and information are widely debated in the literature today (Pincus, 2011). The debate is, in part, driven by methodological issues brought to light by researchers

concerned over how quality process measures gathered by agencies like CMS are being used to rate provider performance (Axon & Williams, 2011; Hofer, Hayward & Greenfield, Wagner, Kaplan, & Manning, 1999).

Quality process measures provide information on how often a healthcare provider delivered care that research has linked to positive health outcomes for patients (Rubin, Pronovost, & Diette, 2001). Unlike outcome measures, quality process measures do not require risk adjustment in order to use in comparisons due to the fact all patients, regardless of age, diagnosis, ethnicity, gender, etc. are appropriate for the quality process (e.g., all patients who smoke should be offered information on how to stop smoking) (Iezzoni, 2003; Palmer, 1997). Therefore, comparing providers becomes a percentage calculation given that 100% of eligible patients in the comparison should receive the quality process. Percentages are calculated by dividing the number of patients that received the quality process by the number of patients that were eligible. This percentage can be used to compare providers to each other and/or to predetermined benchmark percentages (e.g., 90% of all patients who smoke will be offered smoking cessation interventions). The simplicity of this approach is appealing but not without its shortcomings.

Healthcare provider reliability refers to the ability of a quality measure to distinguish a healthcare provider performance (i.e., either at the individual provider or organizational level) on a particular quality process measure from the performance of healthcare providers overall. Healthcare provider reliability requires the following factors: (1) a sufficient number of patients eligible for a given quality process measure and (2) performance variation across providers on that quality process measure. The greater the number of patients who are eligible for a quality

process measure, the more precise the estimate of that provider's performance. When performance variation for a given quality process measure across providers is limited, the likelihood that a provider's performance is statistically significantly different from the comparison provider is also decreased. Hofer, Hayward, Greenfield and colleagues (1999) showed that not controlling for provider reliability significantly misrepresented performance differences across providers. Additionally, when the number of patients eligible (i.e., the patient volume) for a quality process is very large or very small for one provider compared to another accurate comparisons become difficult (Fung, Schmittdiel, Fireman, Meer, Thomas, Smider, Hsu, & Sleby, 2010). Authors of numerous studies have found that providers with large patient volumes have better patient outcomes (Chowdhury, Dagash, & Pierro, 2007; Holt, Poloniecki, Loftus, & Thompson, 2007; O'brien, DeLong, & Peterson, 2008). Studies have also shown that increased performance on quality process measures is linked to better outcomes for patients. For example, Jha, Orav, Li, and Epstein (2007) found an inverse relationship between hospital performance on several quality process measures and patient mortality. However, O'brien, DeLong, and Peterson (2008) have shown that providers with large patient volumes were less likely to be identified as top hospitals when compared to providers with smaller patient volumes on several quality process measures. This counter intuitive finding is attributable to the confounding influence of small denominator/patient volumes in the datasets used to calculate the comparisons.

An approach gaining wide acceptance in the healthcare quality improvement field for controlling the confounding influence of small denominators on process measure comparisons is the Achievable Benchmark of Care (ABC) method (Kiefe, Weissman, Allison, Farmer, Weaver,

& Williams, 1998). The ABC method has been used in numerous studies to identify benchmark levels of performance on process measures of care and increase the use of benchmarking procedures by healthcare providers (Allison, Kiefe, & Weissman, 1999; Fukuda, Nakamura, & Takano, 2002; Hinchey et al., 2007; Houston, et al., 2006; Kiefe et al., 2001; Meehan, Stedman, Neuendorf, Francisco, & Neilson, 2007; O'brien, DeLong, & Peterson, 2008; Weissman et al.,1999; Wessell et al., 2008). In response to calls to advance the growing body of knowledge relative to quality improvement through the standardization of terminology and methodology (see Davidoff, 2005; Berwick, 1989 & 2005; Thomson, 2005) the developers of the ABC method describe it as providing "an objective, clinically relevant, data-driven, basis for process of care performance improvement by identifying benchmark care levels already achieved by best-in-class care givers" (http://main.uab.edu/show.asp?durki=14504).

The ABC method can be used to make inter or intra-agency comparisons (i.e., between healthcare organizations or between individual healthcare professionals). In the healthcare quality improvement field benchmarks are typically chosen arbitrarily (e.g., the top ten percent of all providers). This method compares providers to one another by establishing a benchmark that reflects care provided to at least 10% of the total number of patients in the sample space. By limiting the number of patients in the comparison the confounding influence of small denominators is reduced increasing reliability. Wessel and Kiefe (1998) describe the ABC method as lending objectivity and reliability to benchmarks that have been a widely used, but until now, arbitrarily defined tool. Further, they say the ABC method represents an empirically derived attainable level of excellence for providers to be compared (Ibid,1998).

Using this method, top performing providers are defined as those serving the top 10 percent of all patients in the sample. This percentage can change depending on the comparison being conducted. The developers suggest that earlier in a quality improvement cycle a narrower benchmark (e. g., 15%) could be chosen to make explicit early gains while taking into consideration the time required for improvements (Weissman et al., 1999). As the provider improves the benchmark can them be adjusted upward. Whatever the cutoff percentage chosen it is described as the Benchmark Breakpoint (BB) and is calculated by multiplying the sum of all eligible patients (the denominator values) by the given percentage (e.g., 10% = .10). Provider data is then rank ordered after calculating the Adjusted Performance Fraction (APF).

The APF is a Bayesian estimator. The creation of the APF is attributed to work done by Agresti (1996) and essentially reduces the influence of providers with small denominators while leaving providers who have served more patients/larger denominator percentages less affected. The APF calculation example below demonstrates the effect of this calculation.

Small Hospital Sample: 1 eligible patient receives process/1 eligible patient total = 100%

$$\text{APF} = (1+1)/(1+2) = .67 \quad (1)$$

Large Hospital Sample: 45 eligible patients receive process/60 eligible patients total = 75%

$$\text{APF} = (45+1)/(60+2) = .74 \quad (2)$$

As demonstrated here the smaller denominator is reduced by .33 while the larger hospital percentage was only reduced by .01. The influence of the small denominator size assists in reducing the volume effect on the hospital performance comparison.

After rank ordering the dataset based on the APF values and calculating the cumulative value for the denominators in the dataset the Unadjusted Performance Ratio (UPR) for each

provider is calculated. The UPR is calculated by dividing the number of patients that received the quality process by the number of eligible patients. The final two steps are to use the BB to determine the cutoff point and to calculate the arithmetic mean. The arithmetic mean is calculated by summing the denominators and numerators of all cases at or above the BB and dividing the numerator sum by the denominator sum. All providers with an APF equal to or greater than the arithmetic mean are considered at or above the benchmark and therefore the highest performing.

Although the ABC method does provide a more objective benchmarking method for the healthcare field one component may benefit from further refinement. The arithmetic mean calculation within the ABC method is easily influenced by extreme values either large or small (Einsenhart, 1972; Wilcox, 1995). A good indicator of central tendency when the distribution is normal, the mean finite sample breakdown point is 1/n meaning the proportion of large deviations from the center of a distribution need only be greater than zero for the mean to deviate from the center of the distribution. Other measures of central tendency, such as the median with a finite sample breakdown point of approximately 1/2, are more resistant to large deviations.

The lack of resistance of the arithmetic mean is likely to cause the ABC method benchmark to be an unreliable estimate of performance for some skewed distributions. A review of the literature finds that the question of whether the ABC method can be made more resistant by accounting for distribution skew and kurtosis through the replacement of the arithmetic mean with a more robust measure of central tendency has not been investigated.

Purpose of the Study

This study will determine if the ABC method can be made more resistant to extreme values and therefore deviations from central tendency. A core component of the ABC method, the arithmetic mean, will be replaced by several measures of central tendency that have high finite sample breakdown points. These measures are the 5%, 10% and 20% trimmed mean, the 15% Winsorized mean or the one-step Huber estimator.

Research Question

The following research question will be investigated in this study:

Which ABC method (i.e., ABC method using the mean, 5%, 10%, or 20% trimmed mean, 15% Winsorized mean or one-step Huber) provides the best estimate of central tendency when tested using real healthcare process data?

Significant to the Field

The outcomes of this study will inform statistical methodologists, healthcare providers, administrators and policy makers about how to make the ABC method for assessing and reporting process measures more resistant to naturally occurring deviations in datasets and therefore more reliable an estimate of quality. Statistical methodologists will be able to use these findings to further refine the ABC method through targeting other components, for example the Adjusted Performance Fraction, therefore further improving the reliability of the method. Health providers will be in a position to better understand the impact of their care provision. Research using the ABC method to promote peer-to-peer dialog about care provision has shown it is a useful means to this end (Kiefe et al., 2001). The logical extension to this finding is that further

improvement of the method could help only enhance this finding. Finally, a better understanding of the capacity of the ABC method to measure process outcomes will give healthcare administrators and policy makers valuable insight into when the ABC method should be used.

Assumptions

This study is based on the assumption that findings generated from this study using real data are relevant to the healthcare field.

Study Limitations

The study will use a large dataset from the Medicare Hospital Compare website (see http://www.hospitalcompare.hhs.gov/staticpages/help/hospital-resources.aspx) which could call into question the generalizability of the findings because the number of central tendency measures, healthcare process measure dataset sizes and sample distributions is limited. Further, the hospitals that submit data to the Medicare Hospital Compare website do this voluntarily. There are no controls for hospitals submitting data that are skewed in the direction of a positive provider report.

CHAPTER 2: LITERATURE REVIEW

This chapter includes a detailed review of quality process measurement and benchmarking. An example of the ABC method will be provided to establish a context for investigating if the method can be made more robust to extreme values by adding different robust measures of central tendency to the final step in the method. It is in this final step where the arithmetic mean (for the remainder of the article described as the mean) will be replaced for the purposes of this study. A comprehensive explanation of skew and kurtosis will be followed by a review of four measures of central tendency and how each could contribute to making the ABC method more robust to extreme values.

Quality Process Measurement & Benchmarking

The divide between healthcare research and practice has been described as a chiasm due to the average 17 years it takes for an empirically tested healthcare intervention to move from the field of research into routine clinical practice (IOM, 2001). One solution to closing this divide has been to provide practitioners with information on the rate at which they and their peers are adopting an empirically tested quality process or processes. Process measures are widely used in the healthcare field to monitor provider service provision quality. A recent example is the United States Federal Department of Health and Human Services issuing hospital value-based purchasing quality process measure requirements issued as a result of the 2010 Affordable Care Act legislation (see http://www.healthcare.gov/news/factsheets/valuebasedpurchasing04292011b.html). Donabedian (1966/2005) provided a comprehensive definition of a quality care process:

> Another approach to assessment is to examine the process of care itself

> rather than its outcomes. This is justified by the assumption that one is interested not in the power of medical technology to achieve results, but in whether what is now known to be "good" medical care has been applied. Judgments are based on considerations such as the appropriateness, completeness and redundancy of information obtained through clinical history, physical examination and diagnostic tests; justification of diagnosis and therapy; technical competence in the performance of diagnostic and therapeutic procedures, including surgery; evidence of preventive management in health and illness; coordination and continuity of care; acceptability of care to the recipient and so on. This approach requires that a great deal of attention be given to specifying the relevant dimensions, values and standards to be used in assessment. The estimates of quality that one obtains are less stable and less final than those that derive from the measurement of outcomes. They may, however, be more relevant to the question at hand: whether medicine is properly practiced. (p. 694)

Monitoring quality process measures provides information on how often a healthcare provider delivered care that research has linked to positive health outcomes for patients (Rubin, Pronovost, & Diette, 2001). Unlike outcome measures, process measures do not require risk adjustment in order to use in comparisons due to the fact 100% of the patients, regardless of age, diagnosis, ethnicity, gender, etc. are appropriate for the quality process (e.g., all patients who smoke should be offered information on how to stop smoking) (Iezzoni, 2003; Palmer, 1997). Fredericks, Guruge, Sidani, & Wan (2010) found empirical evidence supporting the provision of postoperative cardiac surgery patients with detailed instructions on how to recognize and report the signs and symptoms of medical complications. Failure to provide this quality process was correlated with increases in postoperative hospital readmissions.

Although the standard for all quality process measures is 100%, often organizations will choose an arbitrary number that is higher than their current baseline as a target for improvement. Quality improvement approaches are then employed to achieve the improvement target.

Similarly, the performance of practitioners or organizations that are deemed best in the field or benchmark providers in the delivery of the practice can be used as the target. The performance of these benchmark providers often becomes a standard that others work to replicate (Casey, & Lloyd, 2001). The reason practitioners and organizations look to compare themselves to benchmark providers instead of their own baseline or perfection (i.e., 100%) is because improvement is best achieved through collaborating or learning from what others have done to overcome barriers and improve their practice. This process is known as continuous quality improvement and is accomplished through an iterative process of putting new learning into practice measuring and then comparing the results to the benchmark providers (Besterfield, Besterfield-Michna, Besterfield, & Besterfield-Sacre, 1999).

A difficulty that can emerge when comparing provider performance is the influence of small denominators. Providers who serve different volumes of patients are difficult to reliably compare. For example, if one provider sees only two patients and provides an evidence-based quality process for both he/she has provided the care to 100% of those who were eligible to receive it. When compared with the provider who had 100 patients and provided the same intervention to 95 patients (i.e., 95% performance on the quality process) it becomes easy to see how the discussion moves from which provider should be seen as a benchmark provider to, how does one reliably compare two providers for benchmarking performance?

The Achievable Benchmarks of Care Method

Recognizing the influence of small denominators on quality process benchmarking comparisons, in 1996 the Agency for Healthcare Research and Quality (AHRQ) funded a study

to investigate if an empirically derived benchmarking method could be created so healthcare providers could more reliably compare benchmark performance to one another. The research led to the creation of the Achievable Benchmarks of Care (ABC) method. This method provided empirically derived benchmarks for providers to use when comparing process measure performance while reducing the influence of small denominator effects on provider comparisons (Kiefe et al.,1998). Through reducing the impact of small denominators on provider quality process comparisons the ABC method allowed for a variety of small and large providers to be reliably compared to one another. Since the publication of this method it has been used in numerous studies. The ABC method has been correlated with an increase in the use of quality improvement benchmark data by practitioners (Houston et al., 2006; Kiefe et al., 2001; Wessell et al., 2008, Ornstein et al., 2008; Wessell, Nietert, Jenkins, Nemeth, & Ornstein, 2008). It has been used to benchmark mental health services (Meehan, Stedman, Neuendorf, Francisco, & Nellson, 2007), public health services (Allison, Kiefe, &Weissman 1999; Fukuda, Nakamura, & Takano, 2002), stroke care (Hinchey et al., 2008; Jacobs, Baker, Roychoudhury, Mehta, & Levine 2005), ophthalmology services (Castejón-Cervero, Jiménez-Parras, Fernandez-Arias, Teus-Guezala, 2011), cardiac bypass surgery (Holman et al., 2004) and diabetes care (Nicolucci, 2008, MacLean et al., 2004). The method has also been used to examine the impact of patient volume on practitioner and hospital performance assessment (Hofer et al., 1999; O'Brien, Delong, & Peterson, 2008) and is listed by the AHRQ as an innovation tool to improve healthcare quality and reduce disparities (see http://innovations.ahrq.gov/content.aspx?id=401).

The data required for the ABC method are provider level fractions where the denominator (d) represents the number of eligible patients for the healthcare process. The

numerator (x) represents the number of patients that actually received the healthcare process. So in this example of 28 providers (see Table One) the proportion of patients who were eligible for an intervention (d) and those who actually received the intervention (x) are being compared. The numerator value can be zero or an integer equal to or less than the denominator value. The denominator value must be at least 1. The first step in calculating the ABC Method is to sum the denominators and numerators in the dataset to determine the Benchmark Breakpoint (BB) (see Table Two).

Using the ABC method definition of the top 10% of the eligible patient population in the comparison set (in this example 28 providers serving a total of 31,519 patients) make up the top performers or benchmark providers. Ten percent of the eligible patients is 3,152 so that is the BB for this comparison group. As discussed earlier, other quality process target values may be chosen by a provider (e.g., 5% or 15%) as the BB. The next step is to calculate the Adjusted Performance Fraction (APF). The APF adjusts the denominator and numerator to control for cases with small denominator values that are equal to or close to the numerator value (Agresti, 1990). The APF essentially reduces the influence of providers with small denominators on the benchmark provider estimation while leaving providers who have served more patients/larger denominator percentages less affected.

After rank ordering in descending order the dataset based on the APF values and calculating the cumulative value for the denominators in the dataset the Unadjusted Performance Ratio (UPR) for each case is calculated (see Table Two).

(Table One) ABC Model Case Example		
Case	Denominator (d)	Numerator (x)
1	45	43
2	66	10
3	5,555	3,561
4	25	25
5	3,333	3,315
6	1,515	1,212
7	88	66
8	486	355
9	183	25
10	2,151	1,896
11	25	17
12	2	2
13	58	57
14	684	548
15	5,161	4,554
16	66	56
17	1,816	1,715
18	5	5
19	1,495	1,240
20	3	3
21	2,151	187
22	22	4
23	48	35
24	644	584
25	232	232
26	105	98
27	88	76
28	5,467	4,667
Total	31,519	24,588

The UPR is calculated by dividing each case numerator by each case denominator. The final two steps are to use the BB (i.e., 3,152) to determine the cutoff point and to calculate the mean. All cases that fall at or above the BB are in the top 10% of the dataset and therefore considered top

or benchmark performers. The mean is calculated by summing the denominators and numerators of all cases at or above the BB and dividing the numerator sum by the denominator sum (see Table Three). Two cases with a UPR equal to or greater the mean are considered the highest performing cases (see Table Two).

(Table Two) ABC Model Benchmark Breakpoint Example

Case	Denominator (d)	Numerator (x)	Adjusted Performance Fraction (APF) = (x+1)(d+2)	Unadjusted Performance Ratio (UPR) = (x)/(d)	Cumulative (d)
25	232	232	0.996	1.000	232
5	3,333	3,315	0.994	0.995	3,565 *Benchmark Breakpoint (3,152)*
13	58	57	0.967	0.983	3,623
4	25	25	0.963	1.000	3,648
17	1,816	17.15	0.944	0.944	5,464
1	45	43	0.936	0.956	5,509
26	105	98	0.925	0.933	5,614
24	644	584	0.906	0.907	6,258
15	5,161	4,554	0.882	0.882	11,419
10	2,151	1,896	0.881	0.881	13,570
18	5	5	0.857	1.000	13,575
27	88	76	0.856	0.864	13,663
28	5,467	4,667	0.854	0.854	19,130
16	66	56	0.838	0.848	19,196
19	1,495	1,240	0.829	0.829	20,691
14	684	548	0.800	0.801	21,375
20	3	3	0.800	1.000	21,378
6	1,515	1,212	0.800	0.800	22,893
12	2	2	0.750	1.000	22,895
7	88	66	0.744	0.750	22,983
8	486	355	0.730	0.730	23,469

23	48	35	0.720	0.729	23,517
11	25	17	0.667	0.680	23,542
3	5,555	3,561	0.641	0.641	29,097
22	22	4	0.208	0.182	29,119
2	66	10	0.162	0.152	29,185
9	183	25	0.141	0.137	29,368
21	2,151	187	0.087	0.087	31,519
Total	31,519	24,588			
Benchmark Breakpoint = (Sum d)(.10)	3,151				

(Table Three) ABC Model Mean Calculation Example		
Case	Denominator (d)	Numerator (x)
25	232	232
5	3,333	3,315
Total	3,565	3,547
Arithmetic Mean = (sum x)/(sum d)	0.995	

In this example both of the providers at or above the 10% BB cutoff are considered top or benchmark performers. These high performing providers would be sought after to assist others with their improvement efforts.

A review of the literature found the use of the mean as the only measure of central tendency in the ABC method for calculating the BB cutoff point. The terms central tendency or location are used to indicate the center point in a distribution of numbers. Kiefe et al. (1998) described the mean in the context of the ABC method as the pared mean from the Portuguese word pare for ceiling due to the role it plays in determining the top or benchmark providers, in the final benchmark calculation. Depending on the shape of the distribution the mean, or some

other measure of central tendency, will best describe the center of the distribution hence providing the most reliable benchmark cutoff value. Keife reported her team did investigate the use of other measures of central tendency in the development of the method but did not expand on what her team did or why the mean is used as the only measure of central tendency in the final step of the ABC method (Catarina L. Kiefe personal electronic mail communication, December 10, 2010).

Skew and Kurtosis

The mean has a breakdown point of zero as such even one extreme (i.e., outlier value) in a distribution of observations will cause the mean calculation to shift or skew in the direction of the extreme value (Hampel, 1985). The instability of the mean as a measure of central tendency indicates that if the final calculation of the ABC method occurs using a distribution of provider performance values that are skewed toward one or more extreme values that the benchmark calculation will be skewed toward these values.

It could be argued that the mean is a good measure of central tendency for a method that seeks to determine benchmark provider performance which is by definition an extreme score (i.e., the top performing provider(s)). This argument would make sense if the mean was sensitive only to extreme values in the direction of high performing organizations (i.e., those with the best performance scores). However, because the ABC benchmark breakpoint is not restricted to one tail of the distribution low performing providers will have just as much influence on the mean. Therefore, depending on the distribution of values, the skew could shift the mean and produce a

"masking effect", leaving an accurate description of the distribution masked by extreme values (Ibid, 1985, p. 99).

Based on a review of the literature two questions have not been answered regarding the ABC method. The first is whether the method is made less reliable as an indicator of the BB due to the breakdown point of the mean and if so can it be improved through the use of more resistant measures of central tendency? Resistant measures are those measures, not just measures of central tendency, that are resistant to small changes to many data points or large changes in a few (Wilcox, 1997). Resistant statistics have higher breakdown points than the mean and are also described as robust. Coined by Box in 1953, robust statistics have a "remarkable property of 'robustness' to non-normality" (cited in Stigler, 2010, p. 227). In statistics, the terms skew (derived from the French to escape or avoid) and kurtosis (derived from the Greek word for bulge) can help describe the properties of a distribution of values (Everitt, 2002). Skew and kurtosis are shape parameters of a distribution that provide an overall picture of how a dataset is organized. The skew describes the symmetry, or asymmetry, of the dataset. It conveys how the data points are distributed relative to a central point, location or typical value in the distribution (Ibid, 2002). A left or positively skewed distribution has data points clustered to the left of the center point, in the negative direction on the x-axis, with a longer tail trailing off to the right. Conversely, a right or negatively skewed distribution has data points clustered to the right of the center point, the positive direction on the x-axis, with a longer tail trailing off to the left.

Although an imprecise description from the standpoint of mathematic statistics, in the simplest of terms, unlike the skew which addresses the left or right shift of the data points on the x-axis, kurtosis can be understood as measuring the shift of data points in a distribution up or

down the y-axis. Kurtosis is a measure of how peaked or flat the distribution is compared to the bell or symmetrically shape of the standard normal distribution. Kurtosis provides a good estimate of the thickness or heaviness of the distribution tails. The normal distribution is described as mesokurtic. A distribution with high kurtosis is peaked near the mean and falls off sharply is described as leptokurtic. Conversely, a low kurtosis indicates a flattening near the mean with the most extreme case represented by a uniform distribution is described as platykurtic.

Skew and kurtosis are described as the third and fourth moment in a symmetrical or normal distribution. Where N is the distribution sample size, x_i the value of the *i*-th member of the distribution, $\bar{x}$ the mean of the *i* values, and σ^2 the variance.

$$(3)\text{Skewness} = \frac{1}{\mathrm{N}}\left(\sum_{i}^{\mathrm{N}}(x_i - \bar{x})^3\right)\frac{1}{(\sigma^2)^{3/2}}$$

$$(4)\text{Kurtosis} = \frac{1}{\mathrm{N}}\left(\sum_{i}^{\mathrm{N}}(x_i - \bar{x})^4\right)\frac{1}{(\sigma^2)^2}$$

Bulmer (1979) offered general rules for using skew and kurtosis to describe the shape of a distribution. He suggested that if skew is calculated to be < -1 or > 1 the distribution is highly skewed. Values between -1 and -1/2 or between 1/2 and 1 describe a moderately skewed distribution and a skew of between -1/2 and 1/2 an approximately symmetric distribution. Using the normal distribution (i.e. kurtosis =3) as a reference, a kurtosis of ≈ 3 would be considered mesokurtic. Greater than 3 leptokurtic and < 3 platykurtic.

Determining the shape of a distribution can be a first step when considering which measure of central tendency could best define the center point of the ABC method distribution of benchmark values. However, the degree of skew and kurtosis does not translate directly to which measures of central tendency are more resistant or better suited for describing the center point of a given distribution.

Concerns Regarding Use of Skew and Kurtosis

Huber cautioned against the misuse of the skew and kurtosis when considering extreme values (Huber, 1972). He discouraged taking the distribution under study and dissecting it into extreme and normal value distributions from which skew and kurtosis could be calculated and compared because it cannot be assumed the non-extreme values will be normally distributed as some have suggested (Ferguson, 1961). Von Hipple (2005) surveyed introduction to data analysis textbooks to determine how the most common measures of central tendency (e.g., mean and median) are described in relation to skew. He found that fourteen of the eighteen introduction to data analysis textbooks he reviewed provided the rule of thumb stating that the mean will reside to the right of the median under right skew distribution, and to the left of the median under left skew distribution (Ibid, 2005). A further analysis of these assertions found that it is not uncommon, especially with discrete distributions, for the mean, median, and mode (i.e., the most frequently occurring value in a set of values) to not behave in this way.

Although deviations from this general guidance are less likely to occur with continuous data, especially with the median since it by definition divides the distribution area in half, it is worth being cautious about making simple assumptions about how measures of central tendency

will behave in a given distribution. Wilcox and Keselman (2003) warned against using skew or kurtosis to determine which measure of central tendency to use for estimation or comparison. Primary to their concern is the poor performance of skew and kurtosis in estimating error when compared to the mean (Ibid, 2003). However, they did not offer an alternative. Instead, they stated "We have considered many other diagnostic strategies, all of which have proven to be rather unsatisfactory"(Ibid, p.271).

Similarly, common techniques for transforming data to correct for violations to normality and homoscedasticity like the logarithmic or square root transformation do not necessarily alleviate problems related to extreme values (Erceg-Hurn & Mirosevich, 2008; Kilian, Matschinger, Löeffler, Roick, & Angermeyer, 2002; Wilcox, 1998). Huber (1972) warned that after such a transformation the underlying distribution is only “approximately known” (p. 1059) which only complicates a matter for which M and L-estimates (soon to be discussed) are better suited to address. Erceg-Hurn and Mirosevich (2008) pointed out several additional concerns including the failure of such transformations to restore normality or homoscedasticity, the possibility of the rearrangement of the order of means, and the difficulty of interpreting the transformation results.

Given the lack of satisfactory diagnostic techniques to help determine which measure of central tendency to use, especially when making comparisons between distributions, it would appear to make sense that a simple calculation and diagramming of skew and kurtosis can offer a general assistance for tests such as the ABC method where confidence intervals are not being calculated. An example of how the skew and kurtosis of a distribution can be helpful, however not diagnostic, in describing the signs or symptoms of non-normality of a distribution can be

found in Pol, Pascual, & Vazquez (2006). In the end the best method remains conducting tests to determine which distribution under study is best approximated by various approaches (Wilcox, 1996). This is the intent of this study.

Robust Measures of Central Tendency

Robust estimates of central tendency that could be tested to replace the mean in the ABC method include linear combinations of order statistics (i.e., L-estimator) including the median, trimmed mean, and Winsorized mean as well as the maximum likelihood estimator (i.e., M-estimator) called the one-step Huber. Although these modern procedures do transform the distribution it is done by targeting the portion of the distribution with heavy tails or kurtosis (Wilcox & Keselman, 2003).

The median divides an ordered distribution of values in half and has a breakdown point of approximately 0.5. A breakdown point this high can accommodate up to half of the data points in a distribution being extreme. Unlike the mean, however, the median is based on one value, if the set of numbers are even, or the mean of two values if the set is odd in number, therefore excluding the rest of the dataset.

$$(5)\ \bar{x} \equiv \begin{cases} Y_{\left((N+1)/2\right)} & \text{if N is odd} \\ \frac{1}{2}\left(Y_{N/2} + Y_{1+N/2}\right) & \text{if N is even} \end{cases}$$

A benefit described by Kiefe et al., (1998) of the ABC method is that it includes the performance of all providers in the performance calculation. Use of the median would discard

half of the top tier providers and would run counter to the inclusive assumption of the ABC method. Therefore, the median will not be considered as a viable option for replacing the mean in the ABC method.

The trimmed mean is a measure that allows for the breakdown point to be chosen. This is accomplished by rank ordering the data and removing both ends of the distribution of values by a certain percentage, typical between 10% and 20%. Where x_t=trimmed mean and k= trimming value.

$$(6)\ TM(x_t) = \frac{1}{n-2K}\sum_{i=k+1}^{n-K} x_{(i)}$$

Wilcox and Keselman (2003) recommended a 20% trim as an "excellent choice" (p. 267) for controlling Type 1 error. The percentage cut determines the breakdown point. The amount of trim can be in the form of fractions of observations or integers. Larger datasets can accommodate integer trimming while small samples (e.g., <10) have been found to not be possible (see Sawilowsky, 1990). Similar to the concerns realized through the use of the median, symmetrical trimming can lead to a large loss of useable or non-outlier data because the trim is always symmetrical so both sides of the distribution are affected. Wilcox also cautioned that trimming can be less robust to large proportions of extreme values (1997). However, the trimmed mean may prove to be useful for certain distributions especially those with heavy tails therefore it would be appropriate replacement for the mean in the ABC method.

The Winsorized mean approach attempts to correct for the loss of data found when using the median and trimmed mean approaches. Not unlike the trimmed mean either end of the distribution is affected not by cutting but by replacement. Both ends of the distribution, typically

10% to 25%, are replaced by the highest and lowest values in the distribution. The mean of this transformed distribution is then calculated where X_w = Winsorized mean, n = sample size, and g =replacement proportion.

$$(7)\ X_w = \frac{1}{n}\{(g+1)X_{(g+1)} + X_{(g+2)} + \ldots + X_{(n-g-1)} + (g+1)X_{(n-g)}\}$$

The Winsorized mean, while considered robust, is still vulnerable to strongly skewed distributions (Wilcox & Keselman, 2003). However, given the lack of robustness of the mean the Winsorized mean does offer a robust alternative for use in the ABC method.

For the L-estimators described here there is much discussion in the literature about how much trim is appropriate. Wilcox pointed out that a clear determination about this is not available however it is understood that trimming is better than not trimming (Wilcox, as cited in Sawilowsky, 2002). Similar to the choice related to trimming with the L-estimators, the one-step Huber M-estimator statistic requires judgment regarding how to calibrate the weighting factor (also known as the bending or tuning factor) (Ibid, 2002). Unlike the L-estimators discussed earlier the one-step Huber is able to trim asymmetrically favoring the side of the distribution populated by extreme scores while at the same time maintaining a high breakdown point. This statistic uses a maximum likelihood approach where the median is employed to estimate a distribution parameter. Where Ψ = weighting constant, the inverse cumulative distribution function for a standard normal curve (μ=o, σ=1) = 1.8977 (see Wilcox, 1996, p.147), MAD = Median Absolute Difference , i = individual observation, and n = sample size.

$$(8)\ Huber_{(\Psi 1.28)}\ M\ Estimator = \frac{1.8977(MAD)(i_1 - i_2) + \sum_{i=i_1+1}^{n-i_2} x_i}{n - i_1 - i_2}$$

Several weighting factors (Ψ) are available for use including 1.28 and 1.8977 recommended by Sawilowsky (2002) which corresponds to the 90 percentile in the normal distribution meaning only values at a distance equal to or above the Ψ value are weighted. The one-step Huber appears to be an ideal alternative to the mean in the ABC method due to its inclusiveness, i.e., inclusion of observations, and sensitivity by trimming in areas where extreme values are found.

Summary

A review of the ABC method and related literature revealed the method is in wide use in the healthcare field to compare process measure performance between providers and to identify top performing clinicians or organizations for use as benchmarks. Further investigation into the components of the method found the mean is used as the final calculation in making the benchmark performance determination. Given the vulnerability of the mean to extreme values several approaches to identifying and controlling for this vulnerability were investigated. The use of the skew and kurtosis as instruments for helping to describe the distribution of top performing organizations was proposed. However, use of these shape parameters for anything more than an indicator of the distribution shape was not supported in the literature. Several robust measures of central tendency were described as possible alternatives to the mean. These literature review findings pointed to the utility of studying how the reliability and therefore usefulness of the ABC method could be improved by replacing the mean with one or more robust measures of central tendency namely the trimmed and Winsorized means and one-step Huber.

CHAPTER THREE: METHODOLOGY

The Achievable Benchmark of Care (ABC) statistical method is designed to control for the impact of small denominator influence on the interpretation of organizational or clinician performance on a binary measure of health care service quality (Weissman et al.,1999). The improvement measure is considered binary because the organization or clinician's quality performance is based on the number of procedures (e.g., referral for a test following a positive screening for a disease) the organization or clinician did or did not correctly execute (e.g., out of 20 patients that screened positive 10 received referrals for further testing).

$$(9)\ \text{Quality Calculation } [0,1] = \frac{\text{Total \# of Patients Receiving Recommended Service}}{\text{\# Patients Eligible for the Recommended Service}}$$

The ABC method uses the mean in the final calculation for determining which organization or clinician provider is delivering the highest quality of care. The instability of the mean as a measure of central tendency, i.e., the tendency for the calculation to be skewed in the direction of large or small numbers, is well documented (Wilcox, 1995).This study will investigate if the ABC method can be made more robust by using a technique known as Monte Carlo simulation with real data to compare how the different measures of central tendency perform. In addition to the mean, two L-estimators (i.e., the trimmed and Winsorized means) and an M-estimator (i.e., one-step Huber) will be compared. The trimmed mean will be calculated at the 5%, 10% and 20% levels. The sample sizes used in the simulations will be large enough to justify trimming a fraction of an observation by using a percentage weight instead of an integer value when calculating the 5%, 10% and 20% trim therefore allowing for all cases to be included in the calculation (Sawilsowsky, 2002). As a result, the assumption that all data points are used

in the ABC method calculation will be maintained. The Winsorized mean will be calculated at the 20% level and following the recommendations of Sawilowsky, the weighting constant for the one-step Huber will be $\Psi_{1.28} = 1.8977$ (Ibid). Robustness will be determined based on how narrow the interval estimate is to the actual population/dataset mean (i.e., μ) for the different measures of central tendency. The measure of central tendency with the narrowest interval around the population mean estimate will be considered to have the most precision and therefore robustness.

Study Data & Data Collection Procedures

Publically available, de-identified data for 33 healthcare process measures available for download from the online Medicare Hospital Compare Website (Ibid) will be used to conduct Monte Carlo simulations. The data were submitted to Medicare from hospitals in all 50 states and four U.S. territories in October of 2011 (see tables four and five for descriptions of each measure). Monte Carlo simulation uses repeated sampling to determine the properties of some phenomenon or behavior (Sawilowsky, 2003). In this study each of the 33 healthcare process measures will be grouped by process type resulting in five datasets. These datasets will be used as an independent population from which samples will be drawn and analyzed using the aforementioned measures of central tendency.

Each measure of central tendency will be used to calculate the ABC method benchmark breakpoint (BB) for determining the top performing provider(s) from the data. The measure of central tendency that best replicates the population mean across the process measure datasets will be considered the measure of central tendency with the highest degree of precision for use by the

ABC method. The Monte Carlo simulations will be conducted using the publically available R simulation software developed by the R Development Core Team (2009).

The Monte Carlo technique allows for a dataset to be sampled and re-sampled with replacement. The re-sampling process repeats hundreds or thousands of times depending on the analysis (Sawilowsky & Fahoome, 2003). This will allow for a comparison of each measure of central tendency using the same population/process measure data. Not unlike a matched pair design where the within group variability is controlled for therefore allowing between method effects to be detected (Burton, Altman, Royston, & Holder, 2006). These between method effects will be compared using the distance between the population mean, μ and the estimated location of the population mean x, ε = (μ-x). Failed samples will be rejected and the process repeated. A record of the number of failed samples will be kept as a large number of failures can indicate the method will be difficult to replicate in the field (Ibid, 2006). The number of repetitions of the experiment will be 10,000 for each sample ensuring it is sufficiently large to ensure accuracy of the results. The pseudo-random number generator found in R software passes tests for randomness.

The center point or point estimate of the simulation confidence interval for each measure of central tendency will provide a comparison point for testing the hypotheses. Comparisons will focus on the degree to which each measure of central tendency is able to describe the population mean. The root mean square error estimation will be used to make this determination where:

$$(10)\ RMSE = \sqrt{E\left((\hat{\theta} - \theta)^2\right)}$$

$\hat{\theta}$= sample parameter

θ = parameter

The research question under study for this work asks which ABC method (i.e., ABC method using the mean, 5%, 10% and 20% trimmed mean, 15% Winsorized mean or one-step Huber to determine the benchmark breakpoint) provides the least biased indicator of central tendency for sets of real data? The null hypothesis (i.e., H_0) states the intervals around the sample mean generated by the simulations for each measure of central tendency will not be significantly narrower or wider than one another. The alternative hypotheses (i.e., H_1) are based on the literature review findings that indicate the one-step Huber $_{\psi 1.28}$ will more accurately describe the center point of the distributions understudy when compared to the other measures of central tendency. Therefore the one-step Huber $_{\psi 1.28}$ is predicted to have the narrowest interval around the sample mean. Histograms and tables will be used to describe the point estimate findings and comparisons.

Appropriate Sample Size

The sample distributions created using R software will provide for long number sequences before repetition while making sure subsets of these sequences are independent precluding the need to test for randomness. The sample denominators for each simulation will include provider samples of the size 10, 20, 30, 50, and 100. The sample numerator values will range from zero to the total amount of the denominator (i.e., the numerator value cannot exceed the value of the denominator). A review of the literature found studies employing the ABC method using a wide range of sample sizes including large sample sizes (i.e., >10,000 providers) (see Wessell, Liszka, Nietert et al., 2008) and smaller sample sizes (i.e., <100 providers) (see

Kiefe, Allison, Williams et al., 2001). Similarly the datasets in this study will contain a range of small and large patient sample sizes (i.e., between 37 and 3,927, see Table Four) from which to sample.

(Table Four) Overview of the Medicare Hospital Compare Datasets

	Number of Hospital/Providers in the Dataset	Condition the Process Addressed	Process Measure Code	Process Measure Score Range Interval	Range of Patients Sample Sizes per Measure
1	160	Children's Asthma Process of Care Measures	CAC_1	[0.96,1]	3-695
2	160		CAC_2	[0.80,1]	3-695
3	160		CAC_3	[0,1]	3-694
4	3,688	Heart Attack or Chest Pain Process of Care Measures	AMI_1	[0,1]	1-951
5	3,596		AMI_2	[0,1]	1-1,587
6	3,002		AMI_3	[0,1]	1-289
7	2,810		AMI_4	[0,1]	1-618
8	3,608		AMI_5	[0,1]	1-1,493
9	447		AMI_7a	[0,1]	1-37
10	1,620		AMI_8a	[0,1]	1-175
11	1,043		OP_2	[0,1]	1-56
12	2,964		OP_4	[0,1]	1-918
13	4,239	Heart Failure Process of Care Measures	HF_1	[0,1]	1-2,121
14	4,255		HF_2	[0,1]	1-2,585
15	4,095		HF_3	[0,1]	1-776
16	3,972		HF_4	[0,1]	1-412
17	4,325	Pneumonia Process of Care Measures	PN_2	[0,1]	1-1,057
18	4,201		PN_3b	[0,1]	1-1,142
19	4,254		PN_4	[0,1]	1-442
20	4,291		PN_5c	[0,1]	1-1,176
21	4,301		PN_6	[0,1]	1-756
22	4,226		PN_7	[0,1]	1-363
23	3,227	Surgical Care Improvement Project Process of Care Measures	OP_6	[0,1]	1-1,902
24	3,209		OP_7	[0,1]	1-1,890
25	3,488		SCIP_CARD_2	[0,1]	1-2,504
26	3,761		SCIP_INF_1	[0,1]	1-3,927
27	3,759		SCIP_INF_2	[0,1]	1-3,927
28	3,754		SCIP_INF_3	[0,1]	1-3,912

29	1,235		SCIP_INF_4	[0,1]	1-1,249
30	3,798		SCIP_INF_6	[0,1]	1-7,715
31	3,701		SCIP_INF_9	[0,1]	1-3,530
32	3,736		SCIP_VTE_1	[0,1]	1-3,718
33	3,730		SCIP_VTE_2	[0,1]	1-3,718

Figure One

Statistical Analysis

Research Question	Variables	Statistical Analysis
Which ABC method (i.e., ABC method using the mean, 5%, 10% and 20% trimmed mean, 15% Winsorized mean or one-step Huber to determine the benchmark breakpoint) provides the least biased indicator of central tendency for the 33 real datasets used in the Monte Carlo analysis?	<u>Independent Variables</u> Sampling distributions generated from 33 healthcare process measure datasets for: 1. Mean 2. 5% Trimmed Mean 3. 10% Trimmed Mean 4. 20% Trimmed Mean 5. 15% Winsorized Mean 6. One-step $\text{Huber}_{\psi 1.28}$ <u>Dependent Variable</u> The Robustness of the ABC Method	The distance between the mean of the population (i.e., parameter) and the simulated sample mean (i.e., parameter estimate) will be compared using the root mean square error $RMSE = \sqrt{E\left((\hat{\theta} - \theta)^2\right)}$ calculation.

(Table Five) Medicare Hospital Compare Healthcare Process Measure Descriptions			
Condition		Process Measure Code	Process Measure Description
1	Children's Asthma Process of Care Measures	CAC_1	Children Who Received Reliever Medication While Hospitalized for Asthma

2		CAC_2	Children Who Received Systemic Corticosteroid Medication (oral and IV Medication That Reduces Inflammation and Controls Symptoms) While Hospitalized for Asthma
3		CAC_3	Children and their Caregivers Who Received a Home Management Plan of Care Document While Hospitalized for Asthma
4	Heart Attack or Chest Pain Process of Care Measures	AMI_1	Heart Attack Patients Given Aspirin at Arrival
5		AMI_2	Heart Attack Patients Given Aspirin at Discharge
6		AMI_3	Heart Attack Patients Given ACE Inhibitor or ARB for Left Ventricular Systolic Dysfunction (LVSD)
7		AMI_4	Heart Attack Patients Given Smoking Cessation Advice/Counseling
8		AMI_5	Heart Attack Patients Given Beta Blocker at Discharge
9		AMI_7a	Heart Attack Patients Given Fibrinolytic Medication Within 30 Minutes Of Arrival
10		AMI_8a	Heart Attack Patients Given PCI Within 90 Minutes Of Arrival
11		OP_2	Outpatients with chest pain or possible heart attack who got drugs to break up blood clots within 30 minutes of arrival
12		OP_4	Outpatients with chest pain or possible heart attack who got aspirin within 24 hours of arrival
13	Heart Failure Process of Care Measures	HF_1	Heart Failure Patients Given Discharge Instructions
14		HF_2	Heart Failure Patients Given an Evaluation of Left Ventricular Systolic (LVS) Function
15		HF_3	Heart Failure Patients Given ACE Inhibitor or ARB for Left Ventricular Systolic Dysfunction (LVSD)
16		HF_4	Heart Failure Patients Given Smoking Cessation Advice/Counseling
17	Pneumonia Process of Care Measures	PN_2	Pneumonia Patients Assessed and Given Pneumococcal Vaccination

18		PN_3b	Pneumonia Patients Whose Initial Emergency Room Blood Culture Was Performed Prior To The Administration Of The First Hospital Dose Of Antibiotics
19		PN_4	Pneumonia Patients Given Smoking Cessation Advice/Counseling
20		PN_5c	Pneumonia Patients Given Initial Antibiotic(s) within 6 Hours After Arrival
21		PN_6	Pneumonia Patients Given the Most Appropriate Initial Antibiotic(s)
22		PN_7	Pneumonia Patients Assessed and Given Influenza Vaccination
23	Surgical Care Improvement Project Process of Care Measures	OP_6	Outpatients having surgery who got an antibiotic at the right time - within one hour before surgery
24		OP_7	Outpatients having surgery who got the right kind of antibiotic (higher numbers are better)
25		SCIP_CARD_2	Surgery patients who were taking heart drugs called beta blockers before coming to the hospital, who were kept on the beta blockers during the period just before and after their surgery
26		SCIP_INF_1	Surgery patients who were given an antibiotic at the right time (within one hour before surgery) to help prevent infection
27		SCIP_INF_2	Surgery patients who were given the right kind of antibiotic to help prevent infection
28		SCIP_INF_3	Surgery patients whose preventive antibiotics were stopped at the right time (within 24 hours after surgery)
29		SCIP_INF_4	Heart surgery patients whose blood sugar (blood glucose) is kept under good control in the days right after surgery
30		SCIP_INF_6	Surgery patients needing hair removed from the surgical area before surgery, who had hair removed using a safer method (electric clippers or hair removal cream – not a razor)

31		SCIP_INF_9	Surgery patients whose urinary catheters were removed on the first or second day after surgery.
32		SCIP_VTE_1	Surgery patients whose doctors ordered treatments to prevent blood clots after certain types of surgeries
33		SCIP_VTE_2	Patients who got treatment at the right time (within 24 hours before or after their surgery) to help prevent blood clots after certain types of surgery

CHAPTER 4: RESULTS

Five process of care datasets were used (see Table Five) from the Medicare Hospital Compare website that where sampled with replacement 10,000 times. The provider samples sizes used were 10, 20, 30, 50 and 100. Results using the RMSE percentage (i.e., the average square distance between the sample measure of central tendency and the population measure of central tendency) as a measure of comparison between the six different measures of central tendency revealed no significant increase in robustness using measures of central tendency other than the mean when calculating the benchmark breakpoint of the ABC method (see Table Six). Consequently, based on these findings the null hypothesis that a significant difference between the six measures of central tendency was not rejected. Results show that the mean performed better than the other measures of central tendency across the five datasets.

An analysis of the RMSE values (see Table Six) shows that the mean either tied for the lowest RMSE value or had the lowest value overall in 88% of the trials. The mean outperformed all the other measures of central tendency (i.e., no ties for lowest RMSE value) for 20% of the trials. The 20% Trimmed mean had the second lowest overall RMSE performance outperforming all other measures, including the mean, for 8% of the total trials. The 15% Winsorized mean and 5% Trimmed mean were the third (i.e., 60%) and fourth (i.e., 56%) best performing measures scoring the lowest RMSE value or tying for the lowest RMSE value. Thirty-two percent of all the trials had equal RMSE values for all the measures of central tendency. Sixty percent of all the trials had tied RMSE values for two or more measures of central tendency.

The five distributions were found to be strongly, negatively skewed (see Appendix A) indicating providers demonstrated consistently high success in complying with each process

measure. Therefore each of the measures of central tendency used to calculate the benchmark breakpoints were calculated using a uniform or homogenous set of score values. As sample sizes increased the RMSE values decreased reflecting the improved accuracy of the estimate afforded by the larger sample size.

(Table 6) RMSE Percentages & Population Mean Values for Measures of Central Tendency								
			Mean	Trim. 05%	Trim. 10%	Trim. 20%	Winsor. 15%	Huber $_{\psi 1.28}$
Sample Sizes	10	Children's Asthma Process of Care Measures	2.64%	2.69%	2.75%	2.88%	2.68%	2.75%
	20	Children's Asthma Process of Care Measures	1.13%	1.14%	1.16%	1.20%	1.14%	1.15%
	30	Children's Asthma Process of Care Measures	0.61%	0.61%	0.61%	0.61%	0.62%	0.62%
	50	Children's Asthma Process of Care Measures	0.21%	0.20%	0.19%	0.17%	0.21%	0.19%
	100	Children's Asthma Process of Care Measures	0.01%	0.01%	0.01%	0.00%	0.01%	0.00%
Children's Asthma Process of Care Measures Population Value			99.00%	99.00%	99.00%	99.00%	99.00%	99.00%
			Mean	Trim. 05%	Trim. 10%	Trim. 20%	Winsor. 15%	Huber $_{\psi 1.28}$
Sample Sizes	10	Heart Attack or Chest Pain Process of Care Measures	0.17%	0.17%	0.17%	0.18%	0.17%	0.17%
	20	Heart Attack or Chest Pain Process of Care Measures	0.03%	0.04%	0.04%	0.04%	0.03%	0.04%
	30	Heart Attack or Chest Pain Process of Care Measures	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
	50	Heart Attack or Chest Pain Process of Care Measures	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
	100	Heart Attack or Chest Pain Process of Care Measures	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Heart Attack or Chest Pain Process of Care Measures Population Value			99.00%	99.00%	99.00%	99.00%	99.00%	99.00%
			Mean	Trim. 05%	Trim. 10%	Trim. 20%	Winsor. 15%	Huber $_{\psi 1.28}$
Sample Sizes	10	Heart Failure Process of Care Measures	0.63%	0.65%	0.66%	0.69%	0.64%	0.66%
	20	Heart Failure Process of Care Measures	0.23%	0.23%	0.23%	0.25%	0.23%	0.23%
	30	Heart Failure Process of Care Measures	0.12%	0.12%	0.13%	0.13%	0.12%	0.13%
	50	Heart Failure Process of Care Measures	0.05%	0.05%	0.05%	0.05%	0.05%	0.05%

	100	Heart Failure Process of Care Measures	0.01%	0.01%	0.01%	0.01%	0.01%	0.01%
Heart Failure Process of Care Measures Population Value			99.00%	99.00%	99.00%	99.00%	99.00%	99.00%
			Mean	Trim. 05%	Trim. 10%	Trim. 20%	Winsor. 15%	$Huber_{\psi 1.28}$
Sample Sizes	10	Pneumonia Process of Care Measures	0.81%	0.82%	0.83%	0.86%	0.81%	0.83%
	20	Pneumonia Process of Care Measures	0.47%	0.48%	0.49%	0.52%	0.48%	0.49%
	30	Pneumonia Process of Care Measures	0.35%	0.35%	0.36%	0.38%	0.35%	0.36%
	50	Pneumonia Process of Care Measures	0.23%	0.23%	0.23%	0.23%	0.23%	0.23%
	100	Pneumonia Process of Care Measures	0.12%	0.12%	0.11%	0.10%	0.12%	0.11%
Pneumonia Process of Care Measures Population Value			99.00%	99.00%	99.00%	99.00%	99.00%	99.00%
			Mean	Trim. 05%	Trim. 10%	Trim. 20%	Winsor. 15%	$Huber_{\psi 1.28}$
Sample Sizes	10	Surgical Care Improvement Process of Care Measures	0.54%	0.55%	0.56%	0.59%	0.55%	0.56%
	20	Surgical Care Improvement Process of Care Measures	0.28%	0.29%	0.30%	0.32%	0.29%	0.30%
	30	Surgical Care Improvement Process of Care Measures	0.18%	0.18%	0.19%	0.20%	0.18%	0.19%
	50	Surgical Care Improvement Process of Care Measures	0.08%	0.08%	0.08%	0.08%	0.08%	0.08%
	100	Surgical Care Improvement Process of Care Measures	0.02%	0.02%	0.02%	0.02%	0.02%	0.02%
Surgical Care Improvement Process of Care Measures Population Value			99.00%	99.00%	99.00%	99.00%	99.00%	99.00%

CHAPTER FIVE: DISCUSSION

The robustness of the mean as a measure of central tendency to calculate the ABC benchmark breakpoint has gained support as a result of this study. Real data from a publically available hospital process of care measure data warehouse were used to compare several measures of central tendency. Comparisons using Monte Carlo simulation found that the mean performed as well or better than the 5%, 10% and 20% trimmed mean, 15% Winsorized mean and the one-step Huber $_{\psi 1.28}$ across a variety of process of care measures. Several factors are likely contributors to these findings.

The distributions were consistently negatively skewed indicating high scores by most providers on each process measure (i.e., $\mu = 99\%$). Whether this is an artifact of the self-report nature of these data is an open question. Regardless, the effect of having such highly skewed data combined with the design of the ABC method, which isolates the top ten percent of performing providers, resulted in highly uniform data. Robust measures of central tendency are designed to reduce the impact of outlier data and therefore will not perform differently than non-robust measures, such as the mean, when compared using uniform data values. Other factors that could have contributed to the results include the following. The Winsorized and trimmed mean approaches have been found to be less robust than the mean when departures from the assumption that the distribution tails are under examination (Stigler, 1973). With regard to the one step Huber, Hill and Dixon contend that although the approach trims symmetrically using the psi function the resulting weights will not necessarily be applied symmetrically when data are strongly skewed (1982).

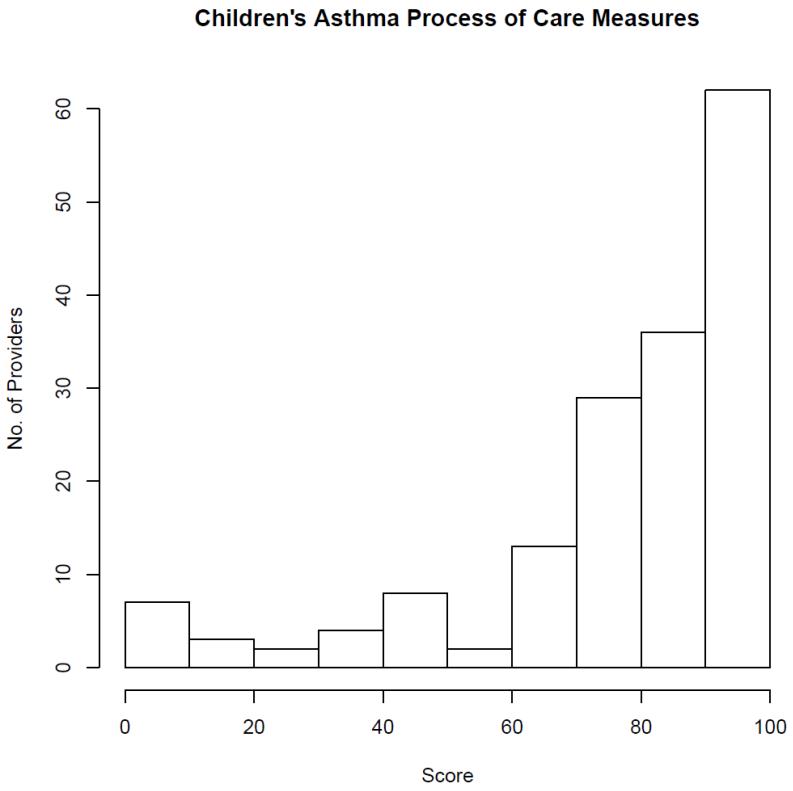
Hill and Dixon (as quoted in Hill & Padmanabhan, 1991) stated, “the general theoretical results about robust estimators do not predict well the true situations...” (p. 81). Rocke, Downs, and Rocke (1982) claimed applied statisticians saw little use for robust statistics and the findings of the Princeton Study of 1972, which were based on small (i.e., < 20), unimodal, and symmetric samples. Stigler (2010) contended that robust estimators work best in the simple analyses with “variations from assumptions that scientists had foreseen” (p. 10), and with the onset of computers and complex statistical analyses that what is meant by robustness is itself becoming more complex. Still Hampel warns, reminded of the importance of understanding the risk of not controlling for nonrobustness, that as massive datasets become ever more accessible to analysis, the inability to draw valid conclusions about the behavior of data remains (2000).

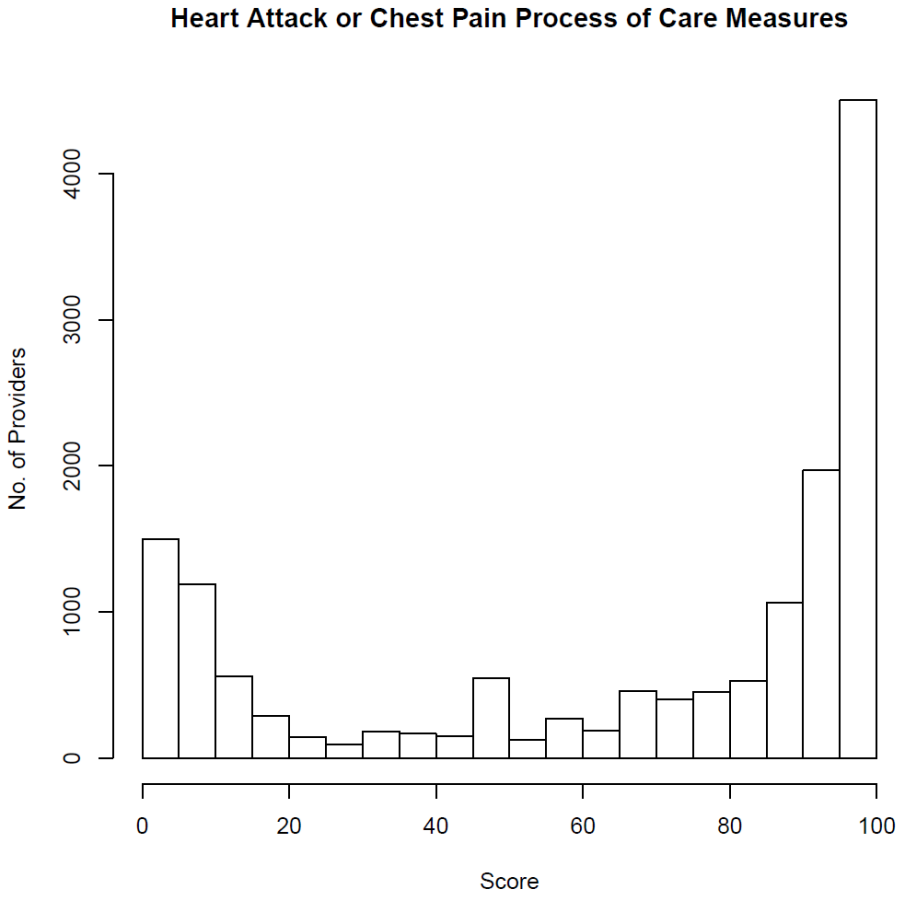
It appears, in many conditions, the ABC method educes a condition where the benchmark breakpoint data are uniformly distributed. It could be suggested the ABC method is a resistant statistic. Regardless, the study findings do not obviate the purpose that drove the development of robust estimators, “exposing more clearly the deviating behavior of parts of the data" (Hampel, 1973, p. 91). Further, robust measures, like those used in this study, allow for down-weighing extreme, yet valid, data leading to more data inclusive and reliable analyses. The question, although partially answered with this study, remains as to whether there are real conditions where the top ten percent of a distribution of process measure scores could contain outlier data. Therefore, creating the condition where the breakdown point of the mean would prove problematic for identifying top performing providers. Simulation studies that test this condition would go a long way in helping to round out the results of this study.

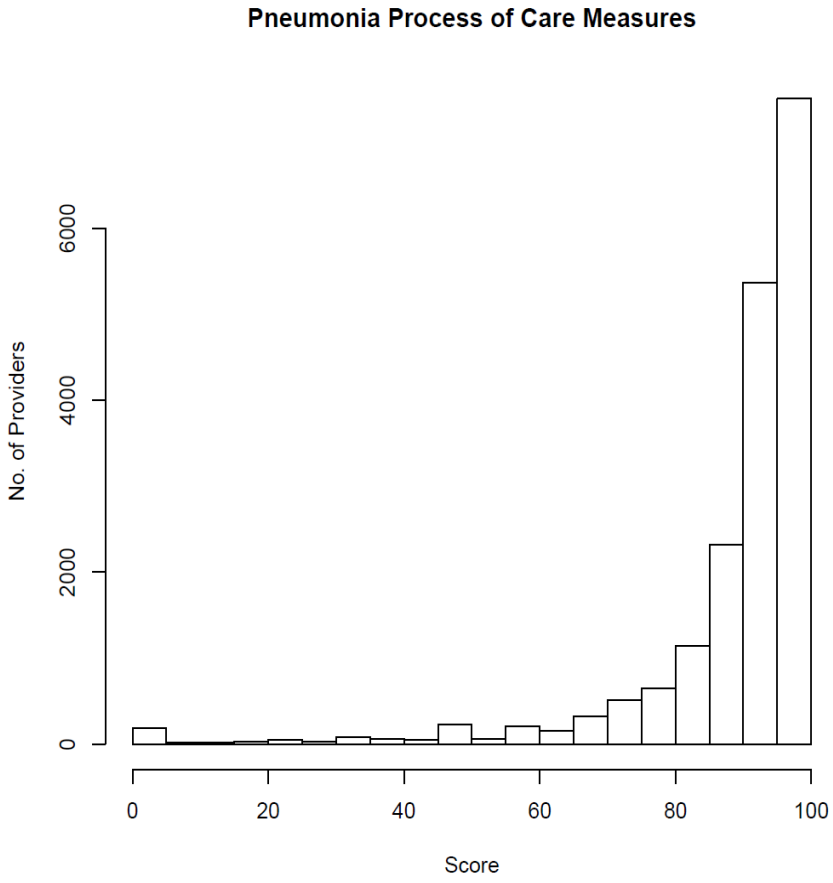
Given these findings healthcare researchers and quality improvement professionals are advised to practice evaluating the distribution shape of large datasets where the ABC method will be used to calculate benchmark providers. For small datasets the examination of each observation is called for where extreme values can be identified and judged to be trustworthy and therefore included or erroneous and therefore excluded from the analysis. If the providers consistently perform well resulting in a distribution that is highly skewed in the direction of high compliance with the process measure this study supports using the mean to calculate the benchmark breakpoint. However, if the data is diagnosed to contain either or both negative and positive outliers (i.e., a variety of low and high performing providers among the benchmark provider group) use of a more robust measure of central tendency than the mean may be warranted given the real data used in this study did not test this condition.

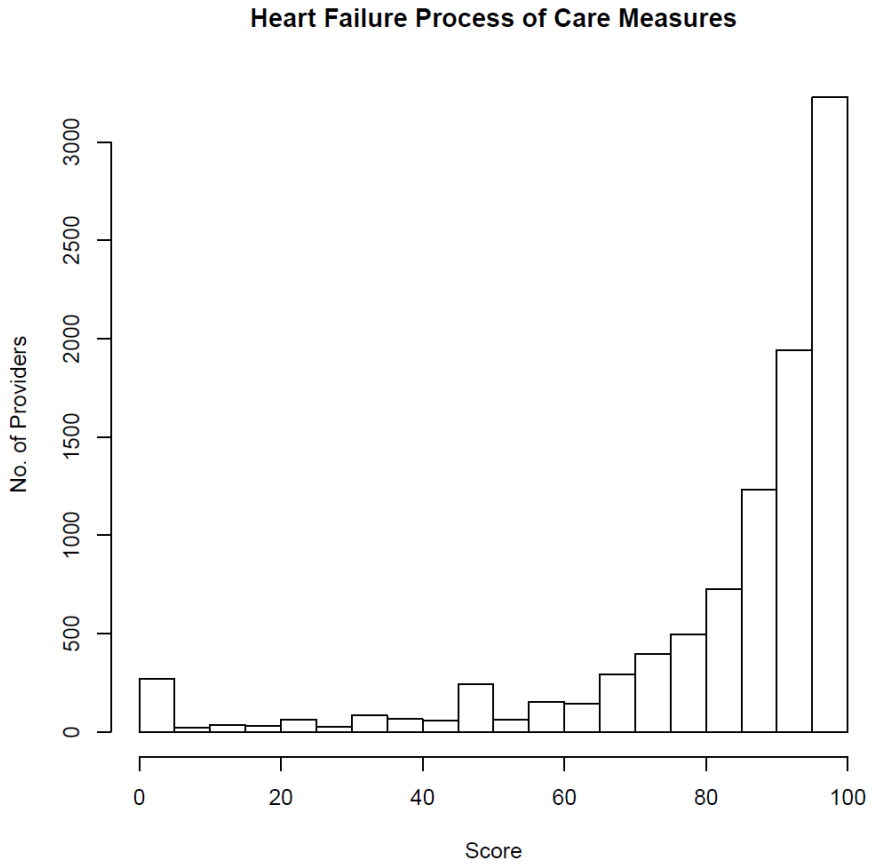
APPENDIX

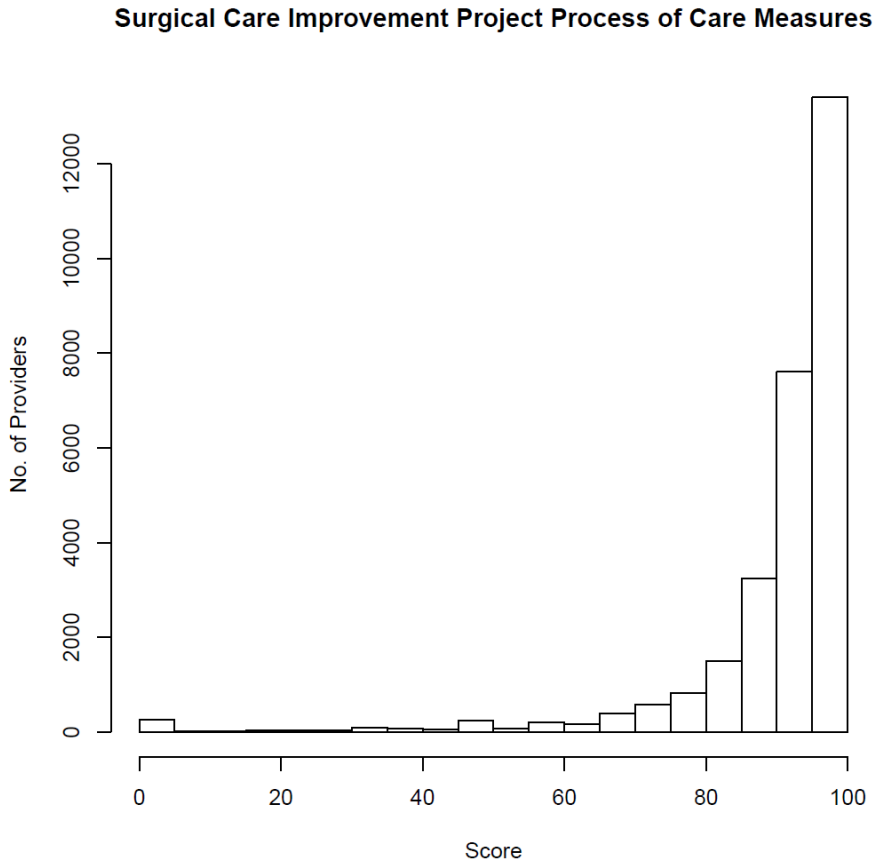
Medicare Hospital Compare Healthcare Process Measure Population Distributions




Heart Attack or Chest Pain Process of Care Measures
No. of Providers
0
1000
2000
3000
4000
0
20
40
60
80
100
Score


Pneumonia Process of Care Measures
No. of Providers
0
2000
4000
6000
0
20
40
60
80
100
Score


Heart Failure Process of Care Measures
No. of Providers
0
500
1000
1500
2000
2500
3000
0
20
40
60
80
100
Score


Surgical Care Improvement Project Process of Care Measures
No. of Providers
0
2000
4000
6000
8000
10000
12000
0
20
40
60
80
100
Score

REFERENCES

Allison, J. J., Kiefe, C. I., & Weissman, N. W. (1999). Can data-driven benchmarks be used to set the goals of healthy people 2010? *American Journal of Public Health, 89*(1), 61-65.

Allison J., Kiefe C. I., Wessman N. W., et al. (1999). *Achievable benchmarks of care (ABC™): User manual. Center for Outcome and Effectiveness Research and Education (COERE).* Birmingham, AL: University of Alabama at Birmingham.

Axon, R. N., & Williams, M. V. (2011). Hospital readmission as an accountability measure. *Journal of the American Medical Association*, *305*(5), 504-505.

Berkey, T. (1994). Benchmarking in health care: Turning challenges into success. *Joint Commission Journal on Quality Improvement, 20*(5), 277-284.

Berwick, D. M. (2005). Broadening the view of evidence-based medicine. *Quality and Safety in Health Care, 14*(5), 315-316.

Besterfield, D. H., Besterfield-Michna, C., Besterfield, G. H., & Besterfield-Sacre, M. (1999). *Total Quality Management* (2nd ed.). Upper Saddle River, New Jersey: Prentice-Hall.

Boscarino, J. A., Chang, J. (1997). Commentary: Inaccurate data on the quality of care may do more harm than good--an alternative approach is required. *American Journal of Medical Quality*, *12*(4), 196-200.

Bulmer, M. G. (1979). *Principles of statistics* (2nd ed.). New York City, NY: Dover Publications.

Burton, A., Altman, D. G., Royston, P., & Holder, R. L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine, 25*(24), 4279-4292.

Carey, R. G. & Lloyd, R. C. (2001). *Measuring quality improvement in healthcare: A guide to statistical process control applications.* Milwaukee, Wisconsin: Quality Press.

Castejón-Cervero, M. A., Jiménez-Parras, R., Fernandez-Arias, I., Teus-Guezala, M. A. (2011). Evaluation of compliance with the EGS guidelines in Spain, using achievable benchmarks of care (ABC™) methodology: The IMCA study. *European Journal of Ophthalmology, 21*(2),149-155.

Curry, L. A., Spatz, E., Cherlin, E., Thompson, J. W., Berg, D., Ting, H. H., ... Bradley, E. H. (2011).What distinguishes top-performing hospitals in acute myocardial infarction mortality rates? A qualitative study. *Annals of Internal Medicine*, *154* (6), 384-392.

Davidoff, F., & Batalden, P. (2005). Toward stronger evidence on quality improvement. Draft publication guidelines: The beginning of a consensus project. *Quality and Safety in Health Care, 14*(5), 319-325.

Donabedian, A. (1966). Evaluating the quality of medical care. *The Milbank Memorial Fund Quarterly, 44*(3), 166-203. (Reprinted in The Milbank Quarterly, Vol. 83, No. 4, 2005, (pp. 691–729) Blackwell Publishing.)

Eisenhart, C. (1971). *The development of the concept of the mean of a set of measurements from antiquity to the present day*. Speech presented at the 131st Annual Meeting of the American Statistical Association. Colorado State University, Fort Collins, Colorado.

Erceg-Hurn, D. M., Mirosevich, W. M. (2008). Modern robust statistical methods: An easy way to maximize the accuracy and power of your research. *American Psychologist, 63*(7), 591-601.

Everitt, B. S. (2002). *The Cambridge dictionary of statistics* (2nd ed.). Cambridge, England: University Press.

Fredericks, S., Guruge, S., Sidani, S., & Wan, T. (2010). Postoperative patient education: A systematic review. *Clinical Nursing Research, 19*(2),144-164.

Fukuda, Y., Nakamura, K., & Takano, T. (2002). A combination of an extrapolation method and benchmark method to develop quantitative health targets for Japan. *Health Policy*, *61*(2), 201-212.

Furguson, T. S. (1961). On the rejection of outliers. *Proceedings of the Berkley Symposium on Mathematical Statistic and Probability, volume 1* (pp. 253-287). Berkeley, CA: University of California Press.

Hampel, F. H. (1973). Robust estimation: A condensed partial survey. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, *27*, 87-104.

Hampel, F. H. (1985). The breakdown points of the mean combined with some rejection rules. *Technometrics, 27*(2), 95-107.

Hampel, F. H. (2000). Robust Inference. *Research Report No. 93. Seminar fur Statistik Eidgenossische Technische Hochschule* (pp. 1-32). CH-8092 Zurich, Switzerland

Hill, M., & Dixon, W.J. (1982). Robustness in real life: A study of clinical laboratory data. *Biometrics, 38*, 377-396.

Hill, N. J., & Padmanabhan, A. E. (1991). Some adaptive robust estimators which work with real data. *Biometrical Journal 33*(1), 81-101.

Hinchey, J. A., Shepard, T., Tonn, S. T., Ruthazer, R., Selker, H. P., & Kent, D. M. (2008). Benchmarks and determinants of adherence to stroke performance measures. *Stroke, 39*(5), 1619-1620.

Hofer, T. P., Bernstein, S. J., Hayward, R. A., DeMonner, S. (1997). Validating quality indicators for hospital care. *Joint Commission Journal on Quality Improvement, 23*(9), 455-467.

Hofer, T. P., Hayward, R. A., Greenfield, S., Wagner, E. H., Kaplan, S. H., & Manning, W. G. (1999). The unreliability of individual physician "report cards" for assessing the costs and quality of care of a chronic disease. *Journal of the American Medical Association*, *281*, (22), 2098-2105.

Holman, W. L., Sansom, M., Kiefe, C. I., Peterson, E. D., Hubbard, S.G., Delong, J.F., & Allman, R. M. (2004). Alabama coronary artery bypass grafting project: Results from phase II of a statewide quality improvement initiative. *Annals of Surgery,* 239, (1), 99-109.

Houston, T. K., Wall, T., Allison, J. J., Palonen, K., Willett, L. L., Kiefe, C. I., ... Heudebert, G. R. (2006) Implementing achievable benchmarks in preventive health: A controlled trial in residency education. *Academic Medicine, 81*(7), 608-616.

Huber, P. J. (1971). Robust statistics: A review. *The Annals of Mathematical Statistics, 43*(4), 1041-1067.

Institute of Medicine, Committee on Health Care in America. (2001). *Crossing the quality chasm: A new health system for the 21st century.* Retrieved from http://www.nap.edu/books/0309072808/html/

Jacobs, B. S., Baker, P. L., Roychoudhury, C., Mehta, R. H., & Levine, S. R., (2005). Improved quality of stroke care for hospitalized medicare beneficiaries in michigan. *Stroke*, 36, 1227-1231.

Kiefe, C. I., Allison, J. J., Williams, O. D., Person, S. D., Weaver, M. T., & Weissman, N. W. (2001). Improving quality improvement using achievable benchmarks for physician feedback. A randomized controlled trial. *Journal of the American Medical Association, 285*(22), 2871-2879.

Kiefe, C. I., Weissman, N. W., Allison, J. J., Farmer, R. M., Weaver, M., & Williams, O. D. (1998). Methodology matters-XII. Identifying achievable benchmarks of care: Concepts and methodology. *International Journal for Quality in Health Care, 10*(5), 443-447.

Kilian R., Matschinger H., Löeffler W., Roick C., & Angermeyer M. C. (2002). A comparison of methods to handle skew distributed cost variables in the analysis of the resource consumption in schizophrenia treatment. *Journal of Mental Health Policy and Economics, 5*(1), 21-31.

Levene, H. (1960). Robust tests for equality of variances. In I. Olkin (Ed.), *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling* (pp. 278–292). Palo Alto, CA: Stanford University Press..

MacLean, C. D., Littenberg, B., Gagnon, M., Reardon, M., Turner, P. D., & Jordan, C. (2004). The vermont diabetes information system: Study design and subject recruitment for a cluster randomized trial of a decision support system in a regional sample of primary care practices. *Clinical Trials*, 1(6), 532–544.

Meehan, T. J., Stedman, T. J., Neuendorf, K. E., Francisco, I. D., & Nellson, M. G. (2007). Benchmarking Australia's mental health services: Is it possible and useful? Australian Health *Review, 31*(4), 623-627.

Mohr, J. J., Mahoney, C. C., Nelson, E. C., Batalden, P. B., & Plume, S. K. (1996). Improving

health care, Part 3: Clinical benchmarking for best patient care. *Joint Commission Journal on Quality Improvement, 22*(9), 599-616.

Neuhauser, D. (2002). Ernest Amory Codman MD. *Quality and Safety in Health Care*, *11* (1), 104-105.

Nicolucci, A. (2008). Five year impact of continuous quality improvement effort implemented by a network of diabetes outpatient clinics. *Diabetes Care, 31*(1), 57-62.

O'Brien, S. M., Delong, E. R., & Peterson, E. D. (2008). Impact of case volume on hospital performance assessment. *Archives of Internal Medicine, 168*(12), 1277-1284.

Ornstein, S., Nietert, P. J., Jenkins, R. G., Wessell, A. M., Nemeth, L. S., Rose, H. L. (2008). Improving the translation of research into primary care practice: Results of a national quality improvement demonstration project. *Joint Commission Journal on Quality and Patient Safety, 34*(7), 379-390.

Palmer, R. H.(1997). Process-based measures of quality: The need for detailed clinical data in large health care databases. *Annals of Internal Medicine, 127*(8), 733-738.

Pol, A. P., Pascual, M. B., & Vazquez, P. C. (2006). Robust estimators and bootstrap confidence intervals applied to tourism spending. *Tourism Management, 27*(1), 42-51

R Development Core Team (2009). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rocke, D. M., Downs, G. W. & Rocke, A. J. (1982). Are robust estimators really necessary? Technometrics, 96-101.

Sawilowsky, S. S. (2002). A measure of relative efficiency for location of a single sample. *Journal of Modern Applied Statistical Methods*, *1*(1), 52-60.

Sawilowsky, S. S. (2003). Invited debate: Target article you think you've got trivials? . *Journal of Modern Applied Statistical Methods*, *2*(1), 218-225.

Sawilowsky, S. S. & Fahoome, G. C. (2003). *Statistics via Monte Carlo Simulation with Fortran*. Rochester Hills, MI: Publisher Journal of Modern Applied Statistical Methods.

Schenker, N. (1985). Qualms about bootstrap confidence intervals. *Journal of the American Statistical Association, 80*(390), 360-361.

Scholle, S. H., Roski, J, Adams, J. L., Dunn, D. L., Kerr, E. A. Dugan, D. P., & Jensen, R. E. (2008). Benchmarking physician performance: Reliability of individual and composite measures. *American Journal of Managed Care, 14*(12), 829–838.

Stigler, S. M. (1973). Asymptotic distribution of the trimmed mean. *The Annals of Statistics, 1*(3), 472-477.

Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900.* Cambridge, MA: The Belknap Press of Harvard University.

Stigler, S. M. (2010). The changing history of robustness. *The American Statistician, 64*(4), 277 281.

Thomson, R. G. (2005). Consensus publication guidelines: The next step in the science of quality improvement? *Quality & Safety in Health Care, 14*(5), 317-318.

Von Hipple, P. T. (2005). Mean, median, and skew: Correcting a textbook rule. *Journal of*

Statistics Education, 13(2). *Retrieved from* www.amstat.org/publications/jse/v13n2/vonhippel.html

Weissmann, N. W., Allison, J. J., Kiefe, C. I., Farmer, R. M., Weaver, M. T., Williams, O. D., … Baker, C. S. (1999). Achievable benchmarks of care: The ABC™s of benchmarking. *Journal of Evaluation in Clinical Practice, 5*(3), 269-281.

Wennberg, J. (1986). Which rate is right? [editorial]. *New England Journal of Medicine*, *314*(5), 310-311.

Wessell, A. M., Nietert, P. J., Jenkins, R. G., Nemeth, L. S., Ornstein, S. M. (2008). Inappropriate medication use in the elderly: Results from a quality improvement project in 99 primary care practices. *The American Journal of Geriatric Pharmacotherapy, 6*(1), 21-27.

Wessell, A. M., Liszka, H. A., Nietert, P. J., Jenkins, R. G., Nemeth, L. S., & Omstein, S. (2008). Achievable benchmarks of care for primary care quality indicators in a practice-based research network. *American Journal of Medical Quality, 23*(1), 39-46.

Wilcox, R. R. (1993). Comparing one-step of location when there are more than two groups. *Psychometrika*, *58*(1), 71-78.

Wilcox, R. R. (1996). *Statistics for the Social Sciences*. San Diego, CA: Academic Press.

Wilcox, R. R. (1998). How many discoveries have been lost by ignoring modern statistical methods? *American Psychologist, 53*(3), 300-314.

Wilcox, R. R., & Keselman, H. J. (2003). Modern robust data analysis methods: Measures of central tendency. *Psychological Methods, 8*(3), 254–274.

Wilcox, R. R., Keselman, H. J., Muska, J., & Cribbie, R. (2000). Repeated measures ANOVA:

Some new results on comparing trimmed means and means. *British Journal Mathematical and Statistical Psychology*, *53*(1), 69-82.

ABSTRACT

ROBUSTNESS OF THE ACHIEVABLE BENCHMARK OF CARE METHOD

by

JEFF ARTHUR CAPOBIANCO

August 2012

Advisor: Shlomo S. Sawilowsky, Ph.D.

Major: Evaluation & Research

Degree: Doctor of Philosophy

The Achievable Benchmark of Care Method is a process of care performance improvement measurement approach for identifying top performing healthcare providers. The purpose of this study was to investigate the robustness of the method. This was achieved by comparing the robustness of the standard ABC method, which uses the mean to calculate the benchmark, to versions of the ABC method where the mean was replaced with either a 5%, 10%, or 20% trimmed mean, a 15% Winsorized mean or the one-step Huber $_{\psi 1.28}$ calculation. Monte Carlo simulations where conducted using publically available, Medicare process of care data. The mean was found to perform as well as or better than the other measures when compared based on the root mean squared error estimate calculation. Cause for these results was found through examination of the sample distributions. Each distribution in the study was strongly, negatively, skewed revealing the benchmark provider comparison data to be uniform.

AUTOBIOGRAPHICAL STATEMENT

Jeff Capobianco has over 20 years of experience working as a clinician, administrator, researcher, and consultant in the field of behavioral healthcare. His interest in research and evaluation evolved out of his experience working as a clinician and a need to better understand how to evaluate the effectiveness and efficacy of behavioral healthcare interventions in order to improve workflow efficiencies and patient outcomes.

Ruffolo, M., & Capobianco, J. (2012). Moving an evidence-based practice into routine mental health care: A multifaceted case example. Social Work in Healthcare. Vol. 51(1), 77-87.

Capobianco, J., Svensson, J., Wiland, S., Fricker, C., & Ruffolo, M. (2008). *Guide to implementing evidence-based practices in mental health.* Rockville, MD: National Council for Community Behavioral Healthcare.

Capobianco, J., & Zimmerman, B. (2010). Leading Healthcare Integration: A change leadership guide for mental health and primary care services integration. Rockville, MD, Publisher National Council for Community Behavioral Healthcare.

Kilbourne, A. M., Irmiter, C., Capobianco, J., Reynolds, K., Milner, K., Barry, K., Blow, F. (2008). Improving integrated general medical and mental health services in community-based practices. *Administration and Policy in Mental Health* 35(5):337-45.

Ruffolo, M., Savas, S., Neal. D., Capobianco, J., & Reynolds, K. (2008). The challenges of implementing an evidence-based practice to meet consumer and family needs in a managed behavioral health care environment. *Social Work and Health Care* (Vol. 6), pp. 30-41.

Capobianco, J. (2007). Examples of effective community services and training in family psychoeducation. In Groggatt, D., Fadden, G., Johnson, D.L., Leggett, M., & Shankar, R. (Eds.) *Family as partners in mental health care: A guidebook for implementing family work.* Toronto, Canada: World Fellowship for Schizophrenia and Allied Disorders.

Reynolds, K., Chesney, B., & Capobianco, J. (2006). A collaborative model for integrated mental and physical health care for the seriously and persistently mentally ill: The washtenaw community health organization. *Family, Systems, & Health,* Vol. 24, pp. 19-27.