From: Jackson Williams
Sent: Tuesday, December 02, 2014 11:55 AM
To: Balovlenkov, Elena (CMS/CCSQ) (Elena.Balovlenkov@cms.hhs.gov); Andress, Joel (CMS/OCSQ) (Joel.Andress@cms.hhs.gov); Pratt, Mary J. (CMS/CCSQ); Goodrich, Kate (CMS/CCSQ)
Subject: Request for Correction of DFC Star Ratings under the Data Quality Act

Attached is CCSQ's courtesy copy of the Request for Correction we filed today under the Data Quality Act challenging the methodology to be used in the DFC Star Ratings.

The Data Quality Act and the HHS Information Quality Guidelines promulgated pursuant to the Act have been in existence for approximately as long as CMS has collected and reported quality measures. Over the past decade, no stakeholder has ever filed a Request for Correction with CMS pertaining to quality measurements. We take no joy in being the first, but we feel that the circumstances surrounding the formulation of this program—the lack of stakeholder input, the failure to conduct cognitive testing, the departures from best practices for quality measure reporting as enumerated by CMS' own experts, and expanding the use of the nationwide tournament format even as consensus acceptance of that practice has collapsed—merit our taking this unfortunate step.

We suspect that you share our concerns that much has been sacrificed in the rush to roll this out quickly. If the timetable for this program was dictated by others in the executive branch, and you have felt inhibited in pushing back against pressure to expedite it, we hope you will view this filing as an opportunity. If the program is delayed through the Data Quality process, you can assure its proponents outside the agency that CMS did everything it could to expedite implementation, but was thwarted by an OIRA requirement.

We therefore hope that you will consent to adjudication by a truly neutral arbiter outside of CCSQ—we would suggest Linda Magno or Niall Brennan as persons with the expertise and stature to render an impartial judgment—and that you will be candid in acknowledging the shortcomings of both the process and product to that arbiter.

We look forward to working with you in pursuit of our shared goal of strengthening the Medicare program.

Jackson Williams
Director of Government Affairs
Dialysis Patient Citizens
1012 14th Street, NW, Suite #905
Washington, DC 20005
866-877-4242

DPC
DIALYSIS PATIENT CITIZENS

Improving Life Through Empowerment

December 1, 2014

Information Quality Officer
Centers for Medicare & Medicaid Services
7500 Security Boulevard
Baltimore, MD 21244-1850

Re: Request for Correction Based Upon Non-Compliance With Information Quality Guidelines: Star Ratings for Dialysis Facility Compare Website

Dialysis Patient Citizens respectfully petitions the Center for Medicare and Medicaid Services for correction of the information described below.

I. Description of the specific material that needs to be corrected.

This petition pertains to the proposed Star Ratings for CMS' Dialysis Facility Compare (DFC) Website. The program, and the methodology for assigning star ratings to dialysis facilities, are described in documents available here: https://dialysisdata.org/sites/default/files/content/FAQs/DFC%20Star%20Ratings%20FAQs.pdf

II. Reasons for believing the information does not comply with OMB and HHS information quality guidelines and is in error, and supporting documentation.

This Petition contends that the proposed Star Ratings do not comply with the Objectivity and Utility requirements of the Health and Human Services Information Quality Guidelines (hereinafter "Guidelines").[1] We do not contend that a five-star rating system for DFC is *per se* violative of Guidelines, but rather that the format and methodology CMS developed for use effective January 1, 2015 is (a) biased and (b) lacks usefulness to consumers. The Guidelines' definitions of Objectivity and Utility are set forth below, with *emphasis added* to highlight the particular concerns relevant to this Petition:

> Objectivity involves a focus on ensuring that information is accurate, reliable and unbiased and that information products are presented in an accurate, clear, complete and *unbiased* manner. Objectivity is achieved by using reliable data sources and sound analytical

[1] HHS Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated to the Public, available at http://aspe.hhs.gov/infoquality/Guidelines/part1.shtml .

1012 14th Street, NW, Suite 1475 • Washington, D.C. 20005 • Toll Free Number 1.866.877.4242 • Fax 1.888.423.5002
www.dialysispatients.org • Email: dpc@dialysispatients. org

techniques, and *carefully reviewing information products prepared by qualified people using proven methods.*

Utility involves the *usefulness of the information to its intended users*. Utility is achieved by staying informed of information needs and developing new data, and information products where appropriate. Based on internal analyses of information requirements, convening and attending conferences, *working with advisory committees and stakeholders*, sponsoring outreach activities, and where appropriate, *testing publications with targeted audiences to ensure relevance, clarity, and comprehensiveness*, HHS agencies keep abreast of information needs.

Below we present three specific reasons why we believe the Star Ratings methodology does not comply with the Guidelines:

A. The DFC Star Ratings do not consider regional disparities in mortality due to socio-demographic factors, and therefore are systematically biased so as to disseminate artificially low scores in places where, in general, mortality is higher than average, and artificially high scores in places where mortality is lower than average.

B. In promulgating the DFC Star Ratings, CMS did not follow processes nor apply performance metrics necessary to ensure the usefulness of the information to its intended users.

C. The DFC Star Ratings as promulgated do not achieve usefulness as that standard is defined for presentation of health care quality information.

There are two specific features of the Star Ratings that we believe detract from its objectivity and utility so greatly as to require correction. The first is the nationwide tournament format in which facilities disadvantaged by reason of their region's low income, history of racial discrimination, or poor population health are in competition with facilities whose patients are not plagued by such problems. The second is the idiosyncratic and non-intuitive methodology for assigning stars, a forced "bell curve" in which top-to-bottom numeric performance rankings are transformed into stars using 30/40/30 cutpoints to distinguish the "average" from above and below average facilities. Such a numeric-to-symbol tranformation is contrary to consumers' reasonable expectations from well-known symbols that are packed with vernacular meaning. We also call attention to the lack of scientific rigor in the process of designing the program.

For the most part, the first contention is addressed to the Objectivity prong of the guidelines and the latter two are addressed to the Utility prong. But as a whole, we believe that the purpose of the guidelines is to ensure a measure of rationality in the collection and presentation of information. In terms of process, the Guidelines envision a system in which peer review, notice to and comment by stakeholders, and cognitive testing of information formats add rigor to an agency's deliberations and promote consensus in its outputs. In terms of substance, the Guidelines seek to achieve authoritative guidance for the public that is empirically based; that is, that decisions are grounded in facts and accepted scientific theory and knowledge, not political priorities or reflexive or rushed execution of agency repertoires.

It is clear that DFC Star Ratings will present "influential scientific information" within the definition of that term for Information Quality purposes. Under the Office of Management and Budget Revised Information Quality Bulletin for Peer Review (April 15, 2004)[2] "each agency shall have a peer review conducted on all influential scientific information that the agency intends to disseminate;" however, it appears that Star Rating methodology was not subject to peer review.

The HHS Guidelines provide that if data and analytic results have been subjected to formal, independent, external peer review, the information may generally be presumed to be of acceptable objectivity. In the absence of peer review, this Petition should be adjudicated without such deference. Further, the methodology was not subject to notice-and-comment rulemaking. As such, the Agency actions challenged here should receive an even higher level of scrutiny under this review.

A. The DFC Star Ratings do not consider regional disparities in mortality due to socio-demographic factors, and therefore are systematically biased so as to disseminate artificially low scores in places where, in general, mortality is higher than average, and artificially high scores in places where mortality is lower than average.

The Guidelines' Objectivity requirements are set forth below, with *emphasis added* to highlight the particular concerns relevant to this Petition:

> "Objectivity" involves two distinct elements, presentation and substance.
>
> "Objectivity" includes whether disseminated information is being presented in an accurate, clear, complete, and unbiased manner. This involves *whether the information is presented within a proper context.*
>
> In addition, "objectivity" involves a focus on ensuring accurate, reliable, and unbiased information. In a scientific, financial or statistical context, the original and supporting data shall be generated, and the analytic results shall be developed, using sound statistical and research methods. If data and analytic results have been subjected to formal, independent, external *peer review*, the information may generally be presumed to be of acceptable objectivity.

It is axiomatic that to be presented in proper context, outcome quality measures must be risk-adjusted to account for different individuals' propensity to suffer complications of care. To date, risk adjustments to CMS quality measurements have been made only for individual patient health characteristics such as age or diagnoses. Further, measurements have been tabulated in a nationwide tournament format, without regard to factors that, outside of CMS' rather insular

[2] http://m.whitehouse.gov/sites/default/files/omb/assets/omb/inforeg/peer_review041404.pdf

mindset, are almost universally considered to have an impact on health outcomes.[3] These factors include:

- Socio-economic factors, including patients' resources such as income and supports in the community, and whether the patient is descended from persons who involuntarily migrated to the United States and were held in servitude.
- Other regional differences, such as the value that different regional subcultures place on health habits.

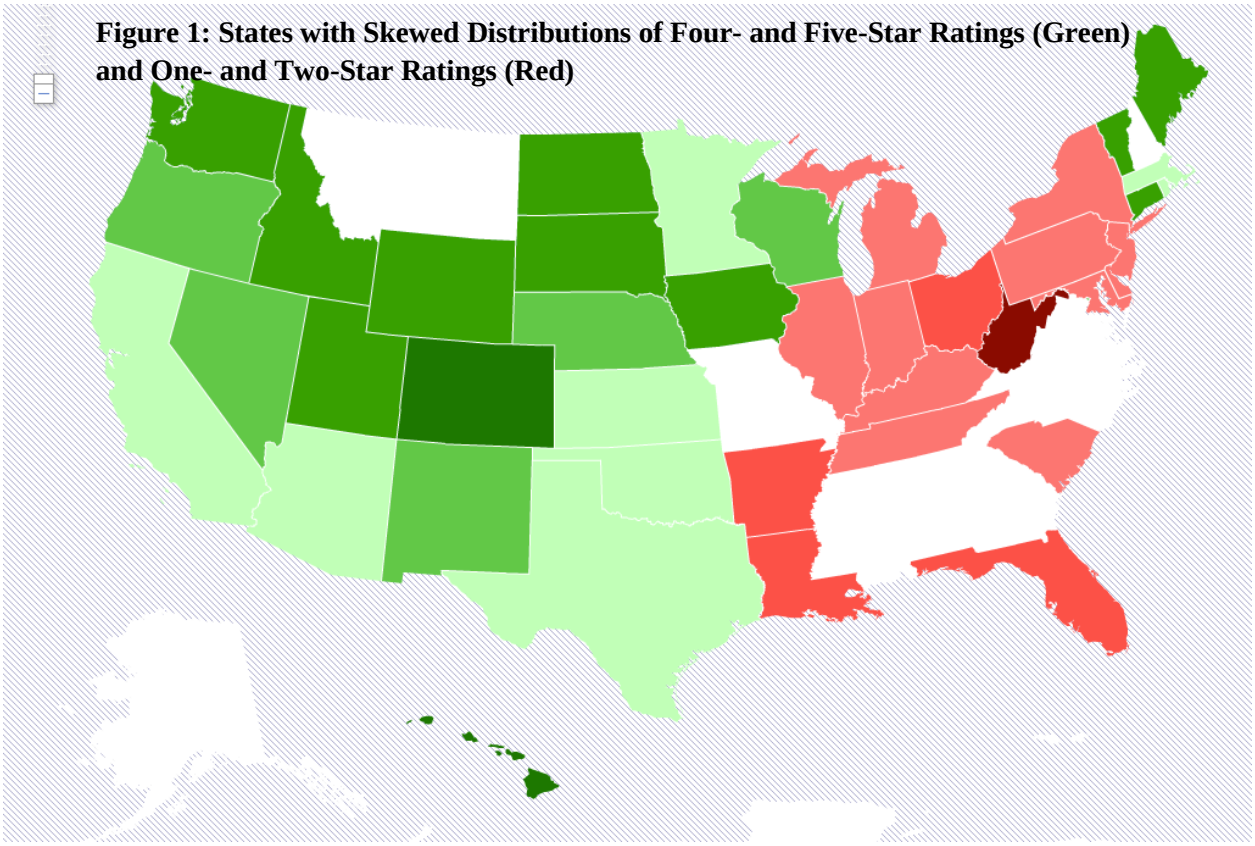
It is apparent that star ratings are not evenly distributed across the country; the Agency has released data indicating that there are a disproportionate number of one- and two-star facilities in the South and Greater Appalachia, and a disproportionate number of four- and five-star facilities in the Pacific Northwest and Upper Midwest.[4] Specifically, all of the top twelve states in percentage of facilities with 4 and 5 stars (three states are tied for 10th at 50%) are located in New England, the West, or Upper Midwest. Of the bottom five states having the highest proportion of 1- and 2-star facilities, four are located south of the Mason-Dixon line. The most extreme disparities on the U.S. mainland are seen in West Virginia, where 63% of facilities are given one- and two-star ratings, and Colorado, where 65% of facilities are given four- and five-star ratings. (Outside the continental U.S., distributions are skewed even more extremely in Puerto Rico at the low end and Hawaii on the high end.)

The map below (Figure 1) summarizes the distribution of stars across states, depicting states having disproportionately more 1- and 2-star facilities in shades of red, and states having disproportionately more 4- and 5-star facilities in shades of green. States having even distribution of stars (that is, 30/40/30) statewide are colored white; however, we do not know if the stars are maldistributed regionally *within* those states. For instance, we suspect that in Virginia, more 1- and 2-star facilities are situated in the Appalachia region, and more 4- and 5-star facilities are located in the affluent D.C. metro area.

[3] See National Quality Forum, *Risk Adjustment for Socioeconomic Status or Other Sociodemographic Factors*; Medicare Payment Advisory Commission, *Report to Congress, June 2013*.

[4] See Appendix 1, "CMS Responses to Questions and Comments about the Dialysis Facility Compare (DFC) Star Rating System."

Figure 1: States with Skewed Distributions of Four- and Five-Star Ratings (Green) and One- and Two-Star Ratings (Red)



The pattern is familiar because it mimics other patterns visible in quality measurements. For instance, in the NCQA annual rankings of health plans released two months ago, all of the top ten commercial plans are located either in New England or Pacific states. Of the top 100 plans, 72 are based in New England, the Upper Midwest, or the Pacific region. Only four are located south of the Mason-Dixon line. Meanwhile, of the bottom 100, 31 are located in the South or Greater Appalachia.

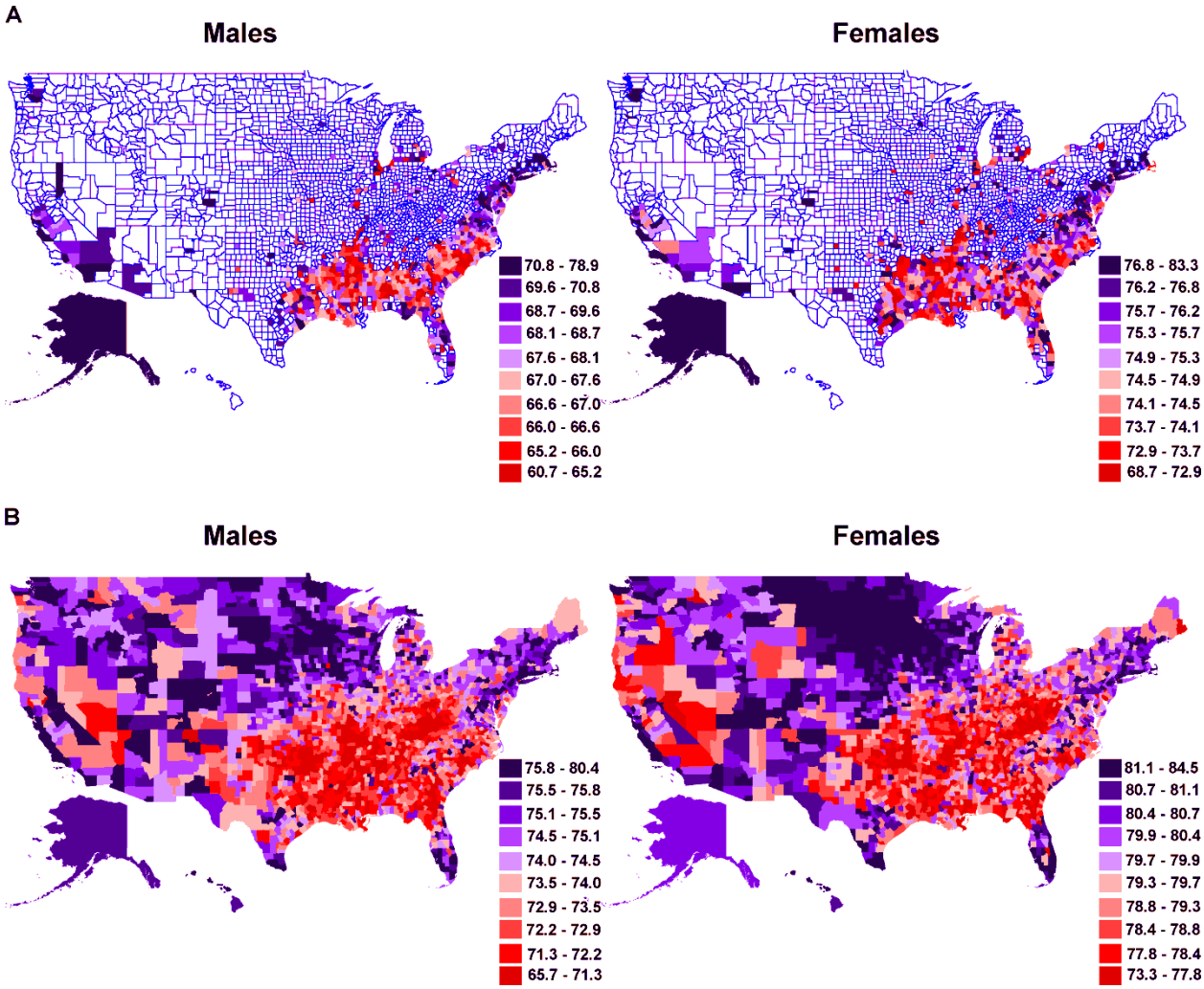
Another example is the hospital readmissions penalty, which assesses most severe penalties against urban safety-net hospitals and hospitals in regions with poor population health. For instance, of 18 hospitals receiving the full 2 percent penalty this past year, ten are in Greater Appalachia, three are in Texas, and two are in Louisiana, with only two north of the Mason-Dixon line.

These patterns closely resemble the maps produced by Christopher Murray and Majid Ezzati that depict life expectancy by county and race to carve out what those researchers dubbed “Eight Americas”—distinct American subpopulations that are either favored or disfavored in terms of health outcomes (see Figure 2).[5] Also apparent are similarities in the Ezzati/Murray maps to a

[5] Murray CJL, Kulkarni SC, Michaud C, Tomijima N, Bulzacchelli MT, et al. (2006) Eight Americas: Investigating Mortality Disparities across Races, Counties, and Race-Counties in the United States. *PLoS Med* 3(9): e260. doi:10.1371/journal.pmed.0030260

map of U.S. Regional Subcultures produced by Joel Lieske. Lieske's work focuses on migrations within the U.S., such as those of Scots-Irish to greater Appalachia, Mormons to Utah and Idaho, and of outdoor and fitness enthusiasts to so-called "Rurban" regions, described as "rural-urban habitats [having] high levels of education, professional and managerial occupations, working women, population mobility, and younger populations… generally found in pastoral academic settings" and particularly in Pacific and Mountain West states.[6]

Figure 2: Life expectancy by county and race (A=Black males and females, B=White males and females)



Taken together, this research shows a pattern in which cultural characteristics of a subpopulation, beyond SES, seem to drive health attitudes and behaviors, leading to differential outcomes.[7] This

[6] Lieske, J. (1993). Regional Subcultures of the United States. *The Journal of Politics*. 55, 888-913.

[7] Williams J. Regional cultures and health outcomes: Implications for performance measurement, public health and policy. *Social Science Journal* 01/2013; 50(4):461–470.

in turn explains why patterns of outcome-based quality measures are wholly predictable and consistent across providers and settings of care.

Comparing the DFC star rating map to the Murray and Lieske maps, we see that West Virginia, Arkansas, and southern Ohio lie in what Murray calls "America 4," an area where life expectancy is "similar to those of Mexico and Panama." (See Appendix 3) Louisiana, South Carolina and Florida lie in what Murray identifies as "America 7," comprised of "low-income rural blacks in the Mississippi Valley and the Deep South," which generally has the highest mortality rates of any of the eight demographic classifications. (See Appendix 4) In terms of regions known for good health, we see that the Dakotas, Iowa and Wisconsin are situated in Murray's "America 2," (See Appendix 3) described as dominated by "low-income rural white populations, with income and education below the national average," emphasizing that SES is a necessary but insufficient adjustment. Star ratings are also high in regions that Joel Lieske classifies as "Rurban," particularly Colorado, Washington State and Oregon.

A recent study found that 58 percent of the variation in hospital readmission rates appears to be associated with where the hospital is located rather than the hospital's performance; that is, the location of the hospital (proximity to other hospitals with high readmission rates) is the best predictor of its readmissions.[8] There is little reason to believe that this same phenomenon is not in play with dialysis facilities.

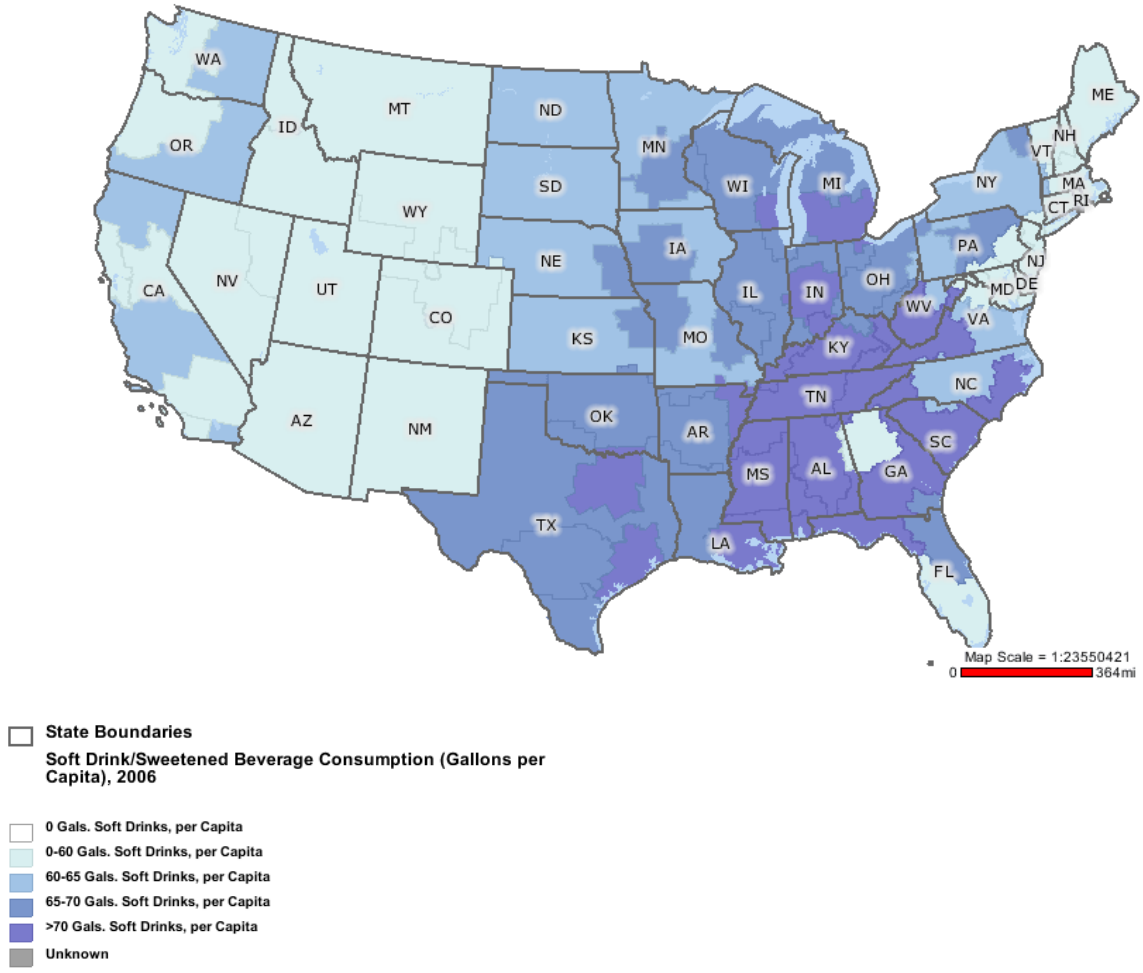
Medicare's risk adjustment techniques do not address these disparities. One obvious reason is the absence of socio-economic status factors in the formula. But even beyond this are regional variations in health behaviors. Consider these differences:

- According to Behavioral Risk Factor Surveillance System (BRFSS) data, recommended fruit and vegetable consumption in Vermont is 50 percent higher than that in West Virginia.[9] The percentage of residents eating five or more servings a day exceeds 28% in every Northern New England state and 25% in every Pacific state, but does not exceed 17% in West Virginia or Louisiana.
- BRFSS data shows that the percentage of people meeting CDC Physical Activity guidelines in West Virginia is less than half that of Colorado. In general the percentage is in the high 20's in the Mountain West but no more than 17 in Greater Appalachia.
- Medication non-adherence in Mississippi is more than double what it is in Wisconsin or New England.[10]
- Soft drink consumption—which can cause complications for ESRD patients—varies by region, and there is a correspondence visible when one juxtaposes the USDA map of per-

[8] Herrin, J., St. Andre, J., Kenward, K., Joshi, M. S., Audet, A.-M. J. and Hines, S. C. (2014), Community Factors and Hospital Readmission Rates. Health Services Research. doi: 10.1111/1475-6773.12177
[9] State-Specific Trends in Fruit and Vegetable Consumption Among Adults --- United States, 2000--2009
[10] CVS Caremark, State of the States: Adherence Report (2012)
http://info.cvscaremark.com/sites/default/files/SOS-Adherence-Report-2013_Final_2.pdf

capita sweetened beverage consumption (Figure 3) with the DFC Star Ratings map. To be sure, we do not assert that there is causal relationship; our concern is that the Agency has not ruled out regional differences prior to insisting upon a nationwide competition.

Figure 3: Per capita Soft Drink Consumption



In Medicare's claims databases, a beneficiary with ESRD in West Virginia can look the same as an ESRD beneficiary in Utah, but we know that the beneficiary in Utah has a completely different profile in terms of health habits and health motivations. We must also be conscious of differences within states. The University of Wisconsin's County Health Rankings website illustrates how dramatically different health indicators can be within a state.

An Inovalon, Inc. study of performance on Medicare Advantage (MA) Star Ratings found that a significant association exists between dual eligible status and lower performance on specific MA

star ratings.[11] The results confirm the integral role that income, and race/ethnicity play on the HEDIS and CMS Part D measures used in the MA Five-Star rating system. The authors' examination of 80 CMS MA contracts found that dual eligible members performed worse on nine of the ten star measures that were investigated, and concluded that that Five-Star rating system penalizes MA plans serving a high proportion of dual eligible beneficiaries.

In sum, because the DFC Star Ratings fail to account for widely-recognized and well-documented socio-demographic factors, the ratings are systematically biased and do not disseminate reliable information about dialysis facilities' quality of care. Instead, they appear to be disseminating information about underlying population health factors in the regions that dialysis facilities serve.

We propounded the following question to the Agency and received the response below (see Appendix 2, "Responses to Questions from Dialysis Patient Citizens"):

> **QUESTION:** If the star ratings were properly risk-adjusted, wouldn't they be evenly distributed across the country without regard to regional population health factors?
>
> **RESPONSE**: CMS has a standing policy not to risk adjust for regional differences in utilization and care.

We believe this type of *ipse dixit* response is precisely what the Guidelines prohibit. The constantly recurring pattern of outcomes for varying types of illness across varying types of providers should have triggered Agency scrutiny of the geographic maldistribution of star ratings prior to their issuance.

We wish to clarify here that this Petition does not challenge the use of nationwide competition in the other quality measures referenced above. We are asking that the nationwide format be replaced in the DFC Star Ratings to be posted in January 2015.

B. In promulgating the DFC Star Ratings, CMS did not follow processes nor apply performance metrics necessary to ensure the usefulness of the information to its intended users.

We begin by noting that CMS' suite of transparency tools was recently evaluated by the Government Accountability Office and found to be deficient. In its October 2014 report, *Health Care Transparency: Actions Needed to Improve Cost and Quality Information for Consumers*, GAO concluded that "CMS has not established procedures and performance metrics to ensure that this information is relevant and understandable to consumers." Further, GAO found that CMS has failed "to develop or select measures that specifically address consumer needs." We

[11] "The Impact of Dual Eligible Populations on CMS Five-Star Quality Measures and Member Outcomes in Medicare Advantage Health Plans." http://www.inovalon.com/inovalon-insights-blog/cms-five-star-quality-rating-system-may-penalize-medicare-advantage-plans

believe that the deficiencies uncovered by the GAO audit prior to development of the DFC Star Ratings were also prominent in the rushing of the un-vetted and untested format and methodology for this program, resulting in a product that fails the Utility prong of the Information Quality Guidelines.

The GAO report recommended that CMS "develop specific procedures and performance metrics to ensure that its transparency tools adequately address the needs of consumers." If procedures and performance metrics incorporating the principles described below had been followed, the Star Ratings would have been made useful to consumers.

1. Both the generally recognized principles of quality reporting and the Information Quality Guidelines emphasize the importance of cognitive testing of the presentation of information to the public prior to its release.

CMS did not conduct research on alternative reporting formats nor conduct cognitive testing of actual measures and displays that are to be put in use. Most importantly, as the Agency stated in its "Responses to Questions from Dialysis Patient Citizens" (see Appendix 2,): "During consumer testing, participants were not explicitly asked for their reactions to a scenario of using a one- or two- star facility. Participants were asked to share their assumptions of the meaning of stars and they overwhelmingly indicated that more stars indicated higher quality." Below, we outline the steps that were skipped in proceeding with the DFC five-star program.

CMS commissioned a report from L&M/Mathematica (hereinafter "L&M") summarizing best practices for presenting consumer information. The report, entitled *Quality Reporting on Medicare's Compare Sites: Lessons Learned from Consumer Research, 2001-2013*, was delivered in September 2014, subsequent to CMS' development of the format and methodology for DFC Star Ratings. In L&M's framework for applying consumer research to quality reporting, the "four P's of marketing" are used as a starting point for discussing consumers' information-seeking and decision-making behavior. Relevant to this discussion is one "P" in particular, price. In the context of usefulness to the public of quality measures, "the price or 'cost' of the product… is related to the cognitive burden of using the information."

As Hibbard & Sofaer explained in a report to AHRQ: "Using quality information to inform choices is hard cognitive work… As the number of pieces of information or decision factors to consider increases, an individual's ability to use that information to make their choices decreases."[12]

For that reason L&M has "used cognitive testing techniques to determine how readily and accurately consumers can understand and interpret narrative content and data displays in

[12] Hibbard J, Sofaer S. *Best practices in public reporting No. 1: how to effectively present health care performance data to consumers*. Rockville, MD: Agency for Healthcare Research and Quality; June 2010. Available at: http://www.ahrq.gov/qual/pubrptguide1.htm.

different formats." Ideally, the process of designing a presentation format for quality measures begins with "a review of the research literature and an environmental scan," followed by development of "communication strategies," then developing and refining products, and finally one-on-one interviews to test "users' understanding and interpretation of information; and usability testing to observe how individuals navigate a particular web tool, what features they use (or miss), and how they accomplish particular tasks."

The need for cognitive testing is heightened by the fact that the DFC Star Ratings are not the only such rating that CMS will present to consumers. Under the ESRD Quality Incentive Program (QIP), CMS is statutorily required to issue each dialysis facility a Total Performance Score to be posted in a prominent place. As the Medicare Payment Advisory Commission (MedPAC) observed in its August 15, 2014 ESRD Comment Letter, "Beneficiaries and their families might be confused if a facility's star and QIP scores diverge… The Commission believes the quality measurement process needs greater simplicity and clarity. Moving to two systems creates greater uncertainty."

Divergence of scores is not a hypothetical scenario. For instance, it appears that the dialysis facility in Hazard, Kentucky—one of the poorest places in the United States—will be given a one- or two-star rating due to its worse-than-expected rates of hospital admissions and mortality. However, this facility has higher than average scores on all four of the clinical measures of quality in the Quality Incentive Program, which means that the certificate in that facility is marked with "Yes" in four boxes under "Meets Standard."

Consumer testing would have revealed how patients react in this scenario. The consumer might appreciate being advised that the QIP gives greater weight to processes taking place within the four walls of the facility while the Star Rating gives more weight to health outcomes. Or the consumer might throw up his hands and feel that contrary scores show that CMS' quality measures lack legitimacy, in which case the DFC Star Ratings will have actually set back the cause of transparency.

Hibbard & Sofaer stress that "While we can do our best to produce reports we think consumers will understand and find meaningful, it is always best to test the information with consumers. Such tests will reveal areas that consumers do not understand, specific misinterpretations, difficulty users have finding information within reports, and users' perception of the information's relevance."

The HHS Information Quality Guidelines call for "testing publications with targeted audiences to ensure relevance, clarity, and comprehensiveness." In addition, the CMS Information Quality Guidelines state: "New and revised information products are tested with focus groups of intended recipients. In many cases, the structure of the content itself is a collaborative process

involving providers, consumers, academicians, and policy analysts."[13] While neither set of guidelines elaborate on the reason for testing, we presume that the intent is the same as that articulated by L&M and Hibbard & Sofaer—to verify the usefulness of the products.

2. CMS failed to follow empirically sound procedures in developing the DFC Star Ratings.

Prior to the CMS Special Open Door Forum conference call in October 2014, the Petitioner propounded the following question to CMS, to which CMS responded as indicated below:

> **QUESTION:** There are a variety of ways to present data to consumers. For instance, both Consumer Reports and NCQA use multiple "bubbles" to represent different individual dimensions instead of using a single composite summary score. So for example, consumer satisfaction is one dimension that both Consumer Reports and NCQA report separately. Can you tell us why the composite five star format, encompassing multiple process and outcome measures, was deemed best suited for dialysis patients?
>
> **RESPONSE**: This was a policy decision. There are a limited numbers of quality measures currently reported on DFC. Multiple component scores are a possibility, but of uncertain value, as the breakdown would probably lead to individual measures receiving their own star rating, and negating the value of providing a summary assessment to patients. (See Appendix 2, "Responses to Questions from Dialysis Patient Citizens".)

CMS' response reveals that the Agency never considered the threshold question of whether a single composite summary score was appropriate for, or the best selection for, the DFC Star Ratings. There was no literature review, no environmental scan, no development of communication strategies, and no testing prior to the selection of the products. Instead, it is quite clear that the decision to use a single score was made by caprice, with minimal consideration of whether consumers found it the most useful medium.

With regard to the final format and methodology that CMS selected, CMS staff has confirmed that the Agency never conducted cognitive testing to determine whether consumers understood what DFC Star Ratings meant and how they differed from other star systems that consumers might be more familiar with.

In analyzing the Medicare "Compare" websites in particular, L&M discussed the different aspects of each provider type and patients' information-seeking and decision-making unique to that transparency tool. L&M concluded "in sum, the circumstances surrounding the kinds of health care decisions Medicare's Compare tools are intended to inform *vary markedly* (emphasis added)." Despite those differences, the Agency has proceeded to implement Star Ratings as if the single composite summary option is equally appropriate for all.

[13] http://aspe.hhs.gov/infoquality/Guidelines/CMS-9-20.shtml

We believe a more sensible model of policy development is that followed by the Federal Highway Administration in revising signage requirements in the Manual of Uniform Traffic Control Devices. This initiative was spurred by concerns that as the average age of motorists increases and their visual acuity decreases, drivers would require street signs that can be more easily read at a distance from a moving vehicle. The agency began this process agnostic about the best medium for presenting this information, but after drawing upon research by ergonomists and human factor specialists on the conspicuity and readability of signs, determined that the size of the letters on street signs should be increased from the current 4 inches to 6 inches, and to phase out capitalized lettering.

Traditionalists have grumbled about this change, but the Department of Transportation can be credited with following the scientific evidence to reach its conclusions. DOT did not decide beforehand that old-style New York City "humpback" street signs were the preferred aesthetic choice, and then tell the public, "you can't make out this sign, so slow down and take a closer look." Such a decision-making process would be analogous to the course CMS followed here: first deciding that a composite star rating would be the most helpful to consumers; then devising its own unique and idiosyncratic system of assigning the stars; and finally attempting to re-educate consumers to substitute the agency's view of what stars mean ("a one-star rating does not mean that you will receive poor care from a facility")[14] for the consumer's own.

C. The DFC Star Ratings as promulgated do not achieve usefulness as that standard is defined for presentation of health care quality information.

The final methodology selected for the DFC Star Ratings ranks every facility in the nation in order on quality scores, and then applies cutpoints at the 90^{th}, 70^{th}, 30^{th} and 10^{th} percentiles to assign stars. As such, the stars give no information on proficiency in any area concrete areas of consumer interest such as those described below. Further, by simply ranking facilities and applying numeric cutpoints, the individual ratings do not have the meanings commonly associated with those symbols in consumers' minds, e.g., that a one-star product or service is to be avoided. As such, the Agency relies on introductory and disclamatory webpage language to re-educate consumers as to the novel meaning CMS wishes the symbols to convey.

Under the Guidelines, "utility refers to the usefulness of the information to its intended users, including the public. In assessing the usefulness of information that the agency disseminates to the public, the agency needs to consider the uses of the information not only from the perspective of the agency but also from the perspective of the public."

We preface this discussion with the caution expressed by L&M: "the context for quality professionals and for consumers is not the same." Conveying the information in a way that

[14] See Appendix 1, "CMS Responses to Questions and Comments about the Dialysis Facility Compare (DFC) Star Rating System."

consumers can understand requires, among other things, "presenting the data in a format that consumers can interpret accurately." While professionals may assume that introductory text, disclaimers, or other explanatory verbiage can fill in gaps, L&M concludes that "most experienced website users do not read introductory text very closely. Instead, they navigate as quickly as possible to what they assume is core content, taking their cues from the organizational features and labels they see. *Categories and terms that make sense to Medicare program personnel or quality professionals may convey something very different to consumers* (emphasis added)."

While there is consensus that summary ratings or composite measures of performance, such as stars or bubbles, are helpful to consumers, L&M warns that there are pitfalls: "consumers often get the wrong idea about what summary ratings actually mean… They tend to interpret such measures in light of what they expect to see… Consumers typically assume, for example, that overall ratings of health care providers or facilities reflect medical experts' judgment" about their quality. This means that the stakes are higher for star ratings than they are for non-summary quality information such as that currently posted to DFC, and the Agency must be extra cautious in ensuring that consumers are not confused or misled. Below we explain why we are convinced that the Agency's failure to perform cognitive testing did in fact result in information that is confusing and misleading.

1. <u>The Star Ratings lack utility because they are counter-intuitive</u>

L&M cautions that "When symbols are not intuitive, these displays place a high cognitive burden on users, who have to locate a legend to determine what the symbols mean and remember that information as they view the data." DFC Star Ratings are not intuitive because they fail to acknowledge consumers' previous experience with star systems. The "bell curve" scoring methodology, which relegates facilities scoring in the bottom three deciles to one- and two-star ratings, does not square with consumers' expectations of what those derogatory symbols connote.

We agree with L&M that "Most consumers are familiar with star ratings from their use in other settings." We note that in the initial announcement of the Star Ratings project by Deputy Director Patrick Conway on the CMS blog, he stated: "Some websites offer 'star' ratings that give information about the quality of the products and services they offer. Wouldn't it be helpful to have the same kind of ratings when choosing a health care provider?"[15]

In referencing existing websites that contain star ratings, Dr. Conway seemed to suggest the purpose of adding stars to DFC was to leverage the public's common understandings of star systems. Therefore, it was a great disappointment to us that the Agency departed drastically from what might be called star system "conventions" or "semiotics," that is, the meanings that are usually ascribed to these five symbols. The Agency has acknowledged the significance of this

[15] http://blog.cms.gov/2014/06/18/star-quality-ratings-coming-soon-to-compare-sites-on-medicare-gov/

departure by proposing to include disclaimer language on DFC reassuring consumers that low star ratings essentially *don't* have the meaning commonly associated with them.

Hibbard & Sofaer note that "Using symbols that are inherently meaningful also can help people quickly discern the meaning of data… The best symbols are those that tell meaning as part of the symbol."

The sphere in which consumers most frequently see star scales is in reviews of discretionary purchases such as movies, restaurants or lodging. In these circumstances, display of the one-star and two-star symbols have an inherent meaning—they are generally understood as advice to *not* make the purchase, e.g. don't see this movie, don't eat at this restaurant. On Amazon.com, a two-star review means "I don't like it" and a one-star review means "I hate it." It is hard to imagine symbols carrying more negative weight.

One application of the "inherent meaning" principle to creation of a systematic star rating program for an industry is the SkyTrax five-star rating system for commercial airlines. SkyTrax rates 190 airlines. One- and two-star reviews are reserved for truly poor performers: only one airline, that of North Korea, receives a single star, and only 23 received two stars.[16] In other words, only the worst twelve percent of all airlines—mostly extreme discount carriers and mostly based in developing countries—received the ratings associated with "do not go."

Until now, CMS has used star rankings for discretionary health care purchases, such as Part C health plans (to which fee-for-service Medicare is an alternative) or nursing homes (to which remaining in the community may be an alternative). If the beneficiary's only options are health plans or nursing homes with one star, he or she can pursue other avenues of receiving care. Dialysis is not a discretionary purchase—it is necessary for a person with kidney failure to stay alive. As such, the stakes involved in assigning low ratings are much higher than they are in other consumer transactions

Below, we have set forth what we believe are the public's common understandings of "inherent meanings" of low star ratings in several ordinary consumer purchase contexts, juxtaposed with the meanings that CMS ascribes to them:

[16] http://www.airlinequality.com/StarRanking/star_system.htm

Common Star Rating Systems	**Dialysis Facility Compare**	**Comment**
In **online product reviews** by buyers, most products receive four or five stars.	In Dialysis Facility Compare, **only 30% of facilities are eligible for four or five stars**.	This was the context noted by Dr. Conway. Studies have found that products reviewed on Amazon.com have a "j-shaped distribution": more positive reviews than negative reviews. This is due to what is known as "purchasing bias": only consumers favorably disposed enough to purchase a product ever review it; and "under-reporting bias": consumers with polarized views are more likely to weigh in.
One or two stars for a **movie** means "thumbs down"– seeing the movie is not a good use of your time and money.	According to CMS, "a one-star rating **does not mean that you will receive poor care** from a facility. It means that the results of one dialysis facility were below average compared to other dialysis facilities."	When there is a need to provide a contrary explanation for something with a widely understood preexisting meaning, it negates the utility of a simplified rating system.
On **Amazon.com**, two stars means *I don't like it* and one star means *I hate it*.	DFC Star Ratings **do not incorporate any patient experience measures.**	We are puzzled as to why Dr. Conway would reference a star rating that can mean the consumer *hates* something, nor a system that uses consumer opinions when DFC will not.
On the website **Yelp,** only 20% of restaurants get one or two stars.	In Dialysis Facility Compare, **30% of facilities automatically get one or two stars**.	Yelp does not use a bell curve, but it limits one- and-two star reviews that are displayed on its website to 20% to counteract under-reporting bias. While one could argue that 30% might be a better ceiling for negative ratings, it is too late to reverse the public's expectations.
In **hotel** classification systems, one or two stars apply to budget lodgings such as motels, with no-frills accommodations, minimal on-site facilities, and no emphasis on comfort and service.	Dialysis Facility Compare star ratings **do not evaluate the physical condition** or amenities of a clinic.	We agree that star ratings should not cover relative comfort or amenities. However, one- and two-star hotels are generally viewed as places one does not patronize unless one cannot afford better or no other accommodation is available.

CMS' justification for departing from the intuitiveness norm relies heavily on the premise that changing symbols' inherent meaning is acceptable when accompanied by disclaimer language. But the Agency has not allowed for what we feel is the high likelihood that that many patients will learn about the star rating for their facility second-hand, without the benefit of such additional information. Our 2014 annual membership survey found that only 11% of our members have ever used DFC, leading us to expect that most patients will learn of a facility's star rating by word-of-mouth.

2. <u>The Star Ratings lack utility because they do not use the type of measures consumers expect.</u>

L&M concluded that "consumers' subjective perceptions of quality [tend] to be based on concrete phenomena they can observe or experience firsthand, rather than abstract value propositions." These factors include technical resources and expertise, paying attention to the patient, management of facilities, operational efficiency, and physical attributes of a facility such as cleanliness. According to L&M, consumers want the data underlying quality measures to reflect meaningful insights into questions such as "'How big a problem is this' 'How much of a risk does this pose' 'How much of a difference is there among providers?'"

DFC Star Ratings fall short because they are not based upon measures that consumers care about, but rather upon the measures that CMS currently has available to choose from. Because CMS has lagged in the collection of dialysis facility CAHPS survey data, there is no patient experience component to the DFC Star Ratings, which, ironically, is the only element to be included in Hospital Compare star ratings at this time. There is no data on physical attributes of a facility, because inspections do not figure in the formula.

Dialysis patients do not connect the complications of illness they experience to their dialysis facility. In our 2014 Annual Patient Survey, administered in July, we asked patients who they thought was in the best position to take steps to avoid their most recent hospitalization.[17] 66 percent said that either the hospitalization was not preventable or that they (the patient) were in the best position. Others felt either their nephrologist or pharmacist was responsible. Only 11 percent felt that their dialysis center was responsible.

As L&M suggests, patients value information about concrete factors: problems or risks they perceive as likely to affect them. Such factors have been overlooked in the rush to post star ratings on DFC.

[17] *Thinking of the most recent episode in which you were hospitalized for your kidney disease, who do you think was in the best position to take steps to avoid that hospitalization?*

3. The Star Ratings lack utility because the cutpoints do not convey information about the magnitude of differences or the degree of variation.

In discussing the use of star ratings, L&M cautions that "they can enhance the impression that ratings are based on expert judgment, as stars are commonly used in other settings (such as hotel ratings and movie or restaurant reviews)… The statistical cutoffs that determine the number of stars displayed may also mislead users about the differences between providers. Also, they do not convey information about the magnitude of differences or the degree of variation."

We believe that in choosing the arbitrary 30/40/30 "bell curve" cutpoints, the Agency has selected the least useful way of distinguishing among providers. Two providers with nearly identical scores can appear, at first glance, to be a quintile apart, simply because they are close to, but on different sides of, a cutpoint. In addition to being untethered to consumers' common understandings of what stars mean, the DFC Star Ratings provide no information on whether a provider meets any absolute standards.

Hibbard and Sofaer inform us that "About one-half of the population has difficulty deriving meaning from numbers. Facing a sea of numbers can be daunting for them, so using symbols rather than numbers can help." But DFC Star Ratings simply translate a precise number into a symbol using an arbitrary formula, only to re-translate the symbol back to a less precise numeric range. This convoluted exercise actually offers less value to the consumer's understanding than would a straightforward ranking of providers, since most patients understand concepts such "class rankings" from their school days. Further, low numeric rankings would have less stigma than low star ratings, since consumers understand that people who graduate at the bottom of the class have still satisfied the requirements of graduation.

We agree with MedPAC that "the measurement of quality performance should be based on absolute standards rather than one calculated from the performance distribution." It is impossible to suppress the suspicion that CMS used the bell-curve methodology in order to sidestep the more complicated task of seeking consensus on absolute standards for dialysis facilities—in other words, that the Agency prioritized a rushed posting of star ratings over deployment of useful star ratings.

4. The Star Ratings lack utility because they use national rather than local reference points

One of the major challenges of presenting quality measures to consumers is impressing upon them how outcomes experienced by other patients are relevant to the quality of care that they are likely to receive. We believe that to be useful, the point of reference must be complications and mortality expected of similarly situated patients. It is not useful for a patient in West Virginia to know what the rates of complications and mortality in nearby facilities are relative to facilities in such far-flung locales as Vermont, Colorado, or Washington State, because the West Virginia patient would have no intention of switching to facilities outside his or her immediate vicinity.

To the extent that the Agency is expecting a West Virginia patient to relocate to a place like Washington State to receive better care, we believe the Agency is contradicting Congress' intention, in enacting the Medicare ESRD benefit in 1972, of ensuring nationwide availability of dialysis facilities and obviating the previous need for kidney failure patients to relocate to receive dialysis.

L&M's research experience echoes this concern: "Many consumers who approach the Compare websites wondering 'what is best for me' are looking to make comparisons in the metropolitan area where they live, and have little interest in benchmarks representing national or state averages."

We wonder what reaction a patient will have to finding that his or her only nearby options for dialysis are facilities with one or two stars. We suspect that this may be the case in those states where CMS has disclosed that a disproportionate share of facilities fall in the bottom 30 percent. We do not have access to the star measures, but we can infer from publicly available information that one such area might be rural Louisiana. Opelousas, Louisiana is a majority African-American town where 43% of the population lives below the poverty line. For a dialysis patient residing in Opelousas, there are eight facilities within 26 miles. Of these eight, three have worse than expected mortality, and only one, located 24 miles away, has a standardized mortality ratio (SMR) below one. The other seven have SMRs of 1.28 or greater. It is not clear to us precisely what our hypothetical Opelousas patient is supposed to do upon being informed that all of his or her nearby options are one- or two-star facilities. It would have been revealing to know how patients reacted to this scenario during cognitive testing, but no such tests were conducted.

For star ratings to be useful, patients in one- and two star facilities must have reasonable access to a three-star or better facility (e.g., a three-star facility within 10 miles or 20 minutes), and patients in "average" facilities must be able to upgrade to an "above average" facility. When a significant number of patients cannot realistically act upon the star ratings, the information is not actionable to the patient and therefore not useful. Further, if low-rated facilities are clustered near one another, and ratings are determined primarily by the region's underlying population health, the ratings are not actionable to facility managers either. To instill a competitive dynamic in which facilities strive for improvement, there must be a realistic opportunity for clinicians to improve outcomes relative to peers. By the same token, we are also concerned that facilities in exceptionally healthy regions, such as New England or the Mountain West, may have artificially inflated rankings solely because their patients are more motivated to adhere to treatment regimens, tempting clinicians to rest on their laurels.

III. Recommendations for correcting the information.

We believe the best way to correct this information is for CMS to follow the process set forth in its own Information Quality Guidelines: "New and revised information products are tested with focus groups of intended recipients. In many cases, the structure of the content itself is a

collaborative process involving providers, consumers, academicians, and policy analysts." This process would solicit stakeholder input on such issues as whether to use absolute standards (and what those standards should be), and whether star ratings for DFC should proceed prior to the availability of patient experience measures; and culminate in cognitive testing of ratings to confirm that dialysis patients understand the specific measures and presentation format.

However, in keeping with those guidelines' requirement that this Petition contain "specific recommendations for correcting the information," we recommend that two principles be incorporated:

A. Use of local reference points for comparisons

There is a great deal of diversity in the United States. This diversity has been recognized by the Medicare program from its outset in contracting with carriers and fiscal intermediaries at the state level, and continues in, for instance, delineating localities for Geographic Practice Cost Indices (GPCIs) or the wage index. The Compare websites themselves recognize the local nature of patient care by grouping providers within a 5, 10, 25, or 50 mile radius of the beneficiary. It therefore strikes us as odd that Mingo County, West Virginia can have a different wage index, a different GPCI, a different Medicare Administrative Contractor, a different Quality Improvement Organization, a different Zone Program Integrity Contractor, and a different local coverage determination for therapeutic shoes, but for quality measurement purposes must share the same expected mortality denominator as every other community from Alaska to Puerto Rico.

We believe that MedPAC has pointed to the correct solution to this problem: instead of holding a national competition, CMS could cluster facilities serving similar patient populations into "peer groups" for quality comparison purposes. In such a regime, the facilities in Louisiana could be judged against each other, not against counterparts in Colorado or Minnesota that set seemingly unattainable standards for them. Clinicians would have to step up their game in every region, because competition would be realistic and the strength of opponents would be no excuse for falling behind. Peer grouping for measures means no reputational punishment for serving disadvantaged communities, so there would be no incentive for national large dialysis organizations to divest facilities in low-income regions to maintain a higher average star rating for their chains as a whole.

We believe the simplest way of accomplishing this would be to devolve comparisons to a meaningful regional unit, which we suspect will not necessarily follow the boundaries of political subdivisions. We think that the Lieske and Murray works cited above offer a roadmap to identifying relatively homogeneous regional clusters, within which facilities would be subject to realistic competition and from which patients could see gradations in ratings that reflect what happens inside the clinic rather than outside the clinic. Localized benchmarks for peer groups of providers that serve relatively homogeneous subsets of patients would engender realistic competition among providers, much as high school size classification systems promote competitive interscholastic sports.

B. Restrict use of symbols to their intuitive and inherent meanings.

As noted above, the one- and two-star symbols have an inherent meaning in the public's mind. A star rating system should meet consumers where they are in their perception of this meaning, and not attempt to indoctrinate website visitors to ignore their experience in, say, rating a product they purchased on Amazon.com with one star because "I hate it." Star ratings for dialysis facilities must also be issued bearing in mind that dialysis, unlike, say, back surgery or angioplasty, is not an elective procedure. One- and two-star ratings are appropriate for hospital departments or specialty practices that can realistically be avoided by prospective patients, because going without treatment from them would not leave the patient worse off.

One- and two-star ratings for dialysis facilities should be limited to those for which CMS has a high degree of certainty that poor outcomes are the result of substandard clinical practices or management, as revealed, for instance, by an inspection, and which a patient should rightfully avoid if an alternative facility is available.

IV. How the Petitioner is affected by the information error.

This Petition is submitted on behalf of Dialysis Patient Citizens (DPC). Dialysis Patient Citizens is America's largest patient-led organization representing dialysis patients, with membership of more than 26,000 dialysis and pre-dialysis patients and their families. DPC's mission is to improve the quality of life of dialysis patients by engaging policy makers, providers and the public. Through patient education, empowerment and advocacy, we work to increase awareness of kidney disease and promote public policy favorable to the widest availability of renal replacement therapy. DPC knows that a diagnosis of ERSD does not mean the end of life. Dialysis patients can lead long and productive lives, but to do so they must maintain a rigorous treatment regimen as well as adhere to strict dietary and fluid intake restrictions.

It is important to the success of star ratings that patients perceive information as actionable and empowering, and not as negative or discouraging. We have two principal concerns about how the Star Ratings system as proposed will impact patients being treated by the 30 percent of providers issued one- and-two star ratings.

First, we are concerned that patients will become unnecessarily alarmed or discouraged about receiving substandard or unsafe care, given the stigma attached to those ratings. Unhappily, these ratings seem likely to prevail in facilities that serve two communities thought to have fatalistic views about illnesses—the African American and Appalachian populations. We acknowledge that there is controversy over whether "fatalism" in these populations is more of a stereotype than a reality; nevertheless, to the degree that it is even a minor problem, policymakers must take care to avoid any possibility of exacerbating it. We do not believe that any gain from the immediate posting of flawed star ratings outweighs the possibility that patients will become discouraged about their prospects of remaining healthy while served by a stigmatized facility.

Second, we fear that the derogatory ratings will undermine patients' trust in the credibility of their clinicians' judgment. Research has demonstrated that patient adherence is influenced by the

credibility patients attribute to members of their care team.[18] Star ratings must not be permitted to reduce patients' trust in their clinicians' knowledge and expertise.

V. Name, mailing address, telephone number, e-mail address, and organizational affiliation, if any, of the individual making the complaint.

This Petition is respectfully submitted by:

Jackson Williams
Director of Government Affairs
Dialysis Patient Citizens
1012 14th Street, NW, Suite #905
Washington, DC 20005
866-877-4242
jwilliams@dialysispatients.org

[18] See, e.g., Lau E et al, Patients' adherence to osteoporosis therapy: exploring the perceptions of postmenopausal women. Can Fam Physician. 2008 Mar;54(3):394-402.