Editorial

Just How Useful Are Health Rankings?

THE PROBLEM WITH RANKS

Ranking health of the counties and states is popular. Such rankings often make the news in local papers, radio, and television news broadcasts. County Health Rankings & Roadmaps (University of Wisconsin Population Health Institute 2014) rankings are done on two components, Health Outcomes and Health Factors. Health Outcomes refer to rates of such things as premature deaths, poor health days, and low birthweight. Health Factors include smoking, obesity, and teen births, as well as other social and economic factors.

According to the 2014 County Health Rankings & Roadmaps (University of Wisconsin Population Health Institute 2014), the county where I live in Iowa (Johnson) has an overall Health Outcomes ranking of 13 out of 99 counties and a Health Factors ranking of 2. However, if I lived a little southeast in Muscatine County, approximately 32 miles by car, things would appear to be quite a bit different. Muscatine County has a Health Outcomes rank of 62 out of 99 and a Health Factors rank of 82.

What are we to do with this information? These ranks can “mobilize” communities to action aimed at addressing the deficiencies. At another level, decision makers might divert programming and funding to shore up the problem areas. The implication is that a low rank suggests there may be a problem with the community’s health that could be made better.

Just how meaningful are these ranks and the differences in the ranks? Previous work has investigated the reliability of rankings in a number of venues (Goldstein and Spiegelhalter 1996; Marshall et al. 1998; Parry et al. 1998; Adab et al. 2002; van Dishoeck et al. 2011; Zhang et al. 2014) and health rankings in particular (Arndt et al. 2011, 2013). Reliability here refers to the degree of instability of the ranks due to random noise. These are just a

few published assessments of the reliability of rankings, but the results appear fairly consistent; that is, rankings can be too unreliable to be helpful.

As an exercise, I calculated the gross overall mortality rate in Iowa as the total number of deaths (Bureau of Vital Statistics, Iowa Department of Public Health 2009) divided by Iowa's 2010 U.S. Census population total. Using this constant mortality rate, I then generated a random Poisson value for each county in Iowa based on the expected number of deaths given the constant state rate and the county population totals. Then, I calculated the counties' observed mortality rate. Note that the rate parameter was constant across all counties, so no county had any more or less "problem" than another. Nonetheless, there was a large difference among the observed rates, randomly affecting county rank. The lowest mortality rate (rank = 1) was in Monroe County with 66.5 deaths per 10,000, and the highest mortality rate (rank = 99) was in Adams County with 114.2 deaths per 10,000. Adams County had over 1.7 times the mortality rate as that of Monroe County. Of course, that is nonsense since these are random variations around a constant rate. Any ranking that I might do on these random mortality rates would be equally random. When I change the seed for the random number generator, a completely different ranking results. The randomness of any policy decisions or funding allocations based on this ranking follows suit.

Even using nonrandom mortality rates, differing counties have different errors of measurement because of the differing population sizes. Does this mean that the ranks in large counties are more meaningful than smaller ones? Unfortunately, ranks are relative to the full set of ranked objects. Consequently, instability in any county affects the rankings of at least a portion of the neighboring ranks. Hence, the problem of unreliability of the entire set of ranks is inherent in the process.

In this issue of *Health Services Research*, Courtemanche, Tchernis, and Soneji recognize these problems. They use a rational and empirically based method to build a set of ranks that improves on the popular County Health Rankings (University of Wisconsin Population Health Institute 2014). There are several innovative aspects to their work.

First, and perhaps most important, Courtemanche, Tchernis, and Soneji use a model that explicitly acknowledges the assumed underlying (i.e., latent)

*Address Correspondence to Stephan Arndt, Ph.D. Department of Psychiatry, Carver College of Medicine; Department of Biostatistics, College of Public Health; Iowa Consortium for Substance Abuse Research, University of Iowa, 100 MTP4 Iowa City, Iowa 52240-5000; e-mail: stephan-arndt@uiowa.edu.

factor, “health.” While this may seem like a small or subtle innovation, it is important. In the end, when we rank a set of objects, we are ranking them on something (e.g., health) that we measured using our indices or indicators. The current paper uses a factor analytic method to tighten up the measure.

Another addition that Courtemanche, Tchernis, and Soneji incorporate is appropriate treatment of missing data. Rather than a static imputation method, these authors use a more reasoned multiple-imputation approach that builds in and accounts for the uncertainty stemming from the missing information. Static imputation errantly ignores this uncertainty and pretends the indicators were observed. While the overall mean might or might not be biased by this approach, the estimated standard error of the estimate is too small. While multiple imputation (MI) is standard fare in many disciplines, this may be MI’s first application in modeling population health rankings.

Finally, the current paper discusses the reliability of the final rankings. Interestingly, this paper finds the reliability of rankings varies markedly from state to state. This finding replicates our previous work using a different method (Arndt et al. 2013). So, while the rankings in Wisconsin may work well and be meaningful, the same rankings might not be as meaningful in a different state. Clearly, that implies that each state needs to consider the reliability of its rankings before making any decisions to act on them.

CONCLUSIONS

Ranking county (or any other small area) health seems simple, intuitive, and useful. The ranks seem easy to interpret, but this appearance can be deceptive. The more important problem comes about when decision makers take action based on the rankings. Rankings are driven, in part, by random noise. Where the noise comes from, survey sampling or random errors around a hard measure (e.g., mortality), makes little difference. The degree to which the rankings are random directly drives the degree to which any decisions based on them are random. The appearance that the decisions are evidence-based confounds and hides the problem. The raw rates and their standard errors are far more transparent.

So, what are we to do? The current issue’s paper offers one solution, to generate more accurate sets of rankings. If nothing else, rankings should be accompanied by some simple measure of their reliability, that is, degree to which they are affected by random noise. While the reliability estimate may

suggest that the rankings are of little use for a state, stakeholders will at least know not to chase after a red herring.

Stephan Arndt

REFERENCES

- Adab, P., A. M. Rouse, M. A. Mohammed, and T. Marshall. 2002. "Performance League Tables: The NHS Deserves Better." *British Medical Journal* 324 (7329): 95–8.
- Arndt, S., L. Acion, K. Caspers, and O. Diallo. 2011. "Assessing Community Variation and Randomness in Public Health Indicators." *Population Health Metrics* 9 (1): 3.
- Arndt, S., L. Acion, K. Caspers, and P. Blood. 2013. "How Reliable Are County and Regional Health Rankings?" *Prevention Science* 14 (5): 497–502.
- Bureau of Vital Statistics, Iowa Department of Public Health. 2009. *2009 Vital Statistics of Iowa*. Des Moines, IA: Iowa Department of Public Health.
- Courtemanche, C., S. Soneji, and R. Tchernis. 2015. "Modeling Area-Level Health Rankings." *Health Services Research* 50 (5): 1413–31.
- van Dishoeck, A.-M., H. F. Lingsma, J. P. Mackenbach, and E. W. Steyerberg. 2011. "Random Variation and Rankability of Hospitals Using Outcome Indicators." *British Medical Journal Quality & Safety* 20 (10): 869–74.
- Goldstein, H., and D. J. Spiegelhalter. 1996. "League Tables and Their Limitations: Statistical Issues in Comparisons of Institutional Performance." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 159: 385–443.
- Marshall, E. C., C. Sanderson, D. J. Spiegelhalter, and M. McKee. 1998. "Reliability of League Tables of *In Vitro* Fertilisation Clinics: Retrospective Analysis of Live Birth Rates Commentary: How Robust Are Rankings? The Implications of Confidence Intervals." *British Medical Journal* 316 (7146): 1701–5.
- Parry, G. J., C. R. Gould, C. J. McCabe, and W. O. Tarnow-Mordi. 1998. "Annual League Tables of Mortality in Neonatal Intensive Care Units: Longitudinal Study." *British Medical Journal*. 316: 1931–35.
- University of Wisconsin Population Health Institute. 2014. "County Health Rankings" [accessed on April 18, 2015]. Available at http://www.countyhealthrankings.org/sites/default/files/state/downloads/CHR2014_IA_v2.pdf
- Zhang, S., J. Luo, L. Zhu, D. G. Stinchcomb, D. Campbell, G. Carter, S. Gilkeson, and E. J. Feuer. 2014. "Confidence Intervals for Ranks of Age-Adjusted Rates across States or Counties." *Statistics in Medicine* 33 (11): 1853–66.