American people enjoy ratings and rankings—athletic teams, restaurants, music albums, television shows, movies, automobiles, books, hotels, cities, and even colleges are featured in annual top 10, top 100, or best of lists. The publications that sponsor such articles command wide appeal among consumers and have spawned a growing rating industry. The rules that underlie such ratings have many features in common: (1) important but poorly defined concepts, like the road handling of a car, are quantified using conveniently available performance metrics (eg, acceleration and braking distances) or subjective surveys of experience; (2) these proxies are then combined into an overall index using arbitrarily chosen decision rules; and (3) the composite index is translated into a ranking list or, more often, a quasi-metric such as a star rating. Not surprisingly, how one selects the individual metrics and constructs decision rules for the composite can dramatically reorder rankings.1

For situations such as sports or cinema, ranking lists (eg, league tables) and star ratings promote friendly jousting and harmless amusement. In other settings, transparency about performance may help consumers choose a superior product and may potentially prompt a complacent business to improve the quality of and user satisfaction with its products.

The rub comes when public policy directs such tactics at complex services such as education or health care, where outcome stakes are higher. When the public is paying for services, it is natural to demand added value for the investment and to request that such value be conveyed in a way that nonexperts can understand. Composites, rankings, and their star-rating derivatives can potentially serve such a purpose. When a choice of health care group is feasible, these shortcuts may promote better decision making by the public than more technical presentations of detailed performance comparisons.2

Yet, for years experts have raised cautionary flags regarding the inherent statistical uncertainties in using league tables or similar approaches.3,4 The policy conundrum becomes further amplified when dubious methods of ordering are used to adjust payment. Health outcomes are demonstrably dependent on underlying health risks determined by genetic, social, educational, economic, and behavioral factors that are beyond the immediate ability of health care professionals to influence. Clinicians who disproportionately serve disadvantaged populations have been shown to score lower on rankings and would therefore be likely to experience a decrease in reimbursement under zero-sum pay-for-performance plans.5 It would be puzzling social policy to shift those dollars toward health care professionals serving the healthy, wealthy, and wise, but that may be the unintended consequence of shortcuts such as composites and star ratings.

An important and timely empirical contribution to this debate is made by Nguyen and colleagues.6 In a cross-sectional analysis, they compared the performance on standard performance measures for diabetes and cardiovascular disease care across 1400 physician groups serving more than 1.6 million enrollees aged 18 to 65 years in a commercial health plan. Performance ratings were adjusted in several ways: (1) age and sex alone (base model); (2) base model plus clinical comorbidities; (3) base model plus area-level sociodemographic characteristics derived from the US census and the American Community Survey data; and (4) full model with base, clinical, and social risk adjustment.

The authors address a critical wrinkle about social risk adjustment—done crudely, it may level the playing field by bulldozing true differences in clinician quality. In a competitive labor market, one
may plausibly speculate that higher-performing clinicians would seek advantageous practice settings, while the lesser-skilled clinicians would be relegated to marginal communities. If such events occurred routinely, social risk adjustment would camouflage true disparities in health care quality and implicitly set a lower standard for clinicians serving socially disadvantaged populations.

To circumvent such confounding, Nguyen and colleagues took special care to distinguish within from between-physician effects by conducting a technically complicated 2-step adjustment process. The first step of the adjustment looked at the association of patient characteristics (clinical and social risk) with performance measure adherence within a clinician group. That was used to calculate an expected performance score. In the second step of the adjustment, the expected performance was compared with observed performance across all clinician groups to create an adjusted performance rating. This elegant solution was built on a similar approach in the same article to analyze social risk factors for hospital readmissions.

The results of this approach raise an apparent paradox—the actual association of social risk adjustment with clinician performance variation is modest overall for intermediate outcomes such as diabetes or cholesterol control, and trivial to nonexistent for process measures. One might hypothesize that more granular data—say, individual rather than area-level social risk factors—would result in even more dramatic adjustment. Yet, despite a modest association with overall performance variance, there was a substantial association with ranking for a number of practices—enough, in practice, to change a star rating or cross the threshold for a payment bonus or penalty.

But this paradox may be more apparent than real, having less to do with the process of social risk adjustment than other limitations of league tables such as the inherent unreliability of most clinical performance measures when applied at the level of an individual health care group. There are many sources of measurement error—statistical imprecision owing to small sample size; coding and diagnostic variances in the electronic health record (which are real but underappreciated); and even laboratory error (eg, glycated hemoglobin determinations can vary ≥10% based on laboratory assay). These sources of “noise” were beyond the scope of this article, but still highlight the fact that fair health care group rankings pose many challenges. Unaccounted-for social risk is only 1 source of potential confusion, and finding the perfect risk adjustment will not necessarily make rankings or star ratings any more fair or useful.

By documenting the significant association of social risk factors with physician group performance comparisons, Nguyen and colleagues have produced an important article that will hopefully provoke rethinking of both health care group scorecards and payment strategies. High-quality empirical evidence such as this analysis is critical to inform national health policy, and the authors have done an important public service.

ARTICLE INFORMATION
Open Access: This is an open access article distributed under the terms of the CC-BY License. © 2019 Francis J. JAMA Network Open.

Corresponding Author: Joseph Francis, MD, MPH, Office of Reporting, Analytics, Performance Improvement, and Deployment (RAPID), Veterans Health Administration, 810 Vermont Ave NW, Washington, DC 20420 (joe.francis@va.gov).

Author Affiliation: Office of Reporting, Analytics, Performance Improvement, and Deployment (RAPID), Veterans Health Administration, Washington, DC.

Conflict of Interest Disclosures: None reported.

Disclaimer: The views expressed in this article are those of the author and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the US government.

REFERENCES


