Commentary on "Comparison of the properties of regression and categorical risk-adjustment models"

Arlene S. Ash[1,2] and Randall P. Ellis[3,2]

January 19, 2016

[1] University of Massachusetts Medical School, Department of Quantitative Health Sciences, Worcester, MA

[2] Verisk Health, Inc., Waltham, MA

[3] Boston University, Department of Economics, Boston, MA

While every risk adjustment model is designed with a specific purpose in mind, and with design specifications that reflect that purpose, risk adjustment models are often applied in settings that differ from the original development environment. Key questions are often: Is model M useful for purpose P? Or, if two models, M and M' are available, which is more useful?

Regardless of its particular structure and the modeling choices made by its developers, how well a model performs for a particular purpose is usually addressed empirically – through quantitative measures, such as how closely its predictions match an outcome of interest. Fuller, et al. (2016) does not consider any quantitative measures. Rather, the essence of its argument is that models with a categorical structure are superior to regression-based models because they are more "clinically meaningful." We disagree.

We base our discussion below on independent studies comparing different risk adjustment frameworks, as well as 30 years of our own experience with risk adjustment. The first author of this paper was the lead developer of a class of models in the 1980s called diagnostic cost group (DCG) models (Ash et al, 1989), which transformed in the 1990s into the hierarchical condition category (HCC) model structure. Early HCC work was led by researchers from Boston University (ourselves and colleagues) and Health Economics Research (now Research Triangle Inc., and notably Gregory Pope) plus a physician team including Lisa Iezzoni, John Ayanian, David Bates and Helen Burstin, all of whom have become well-known health services and health policy researchers in their own right. (Ellis and Ash, 1996, Ash et al, 2000; Pope et al, 2000). The US government uses an HCC-based regression models for risk adjustment: CMS-HCC

models for its Medicare Part C and Part D programs and HHS-HCC models in the individual and small group markets under the Affordable Care Act (Kautter et al., 2014). In addition to these public uses, HCC-based models are widely used in the private sector, originally via a firm DxCG Inc. cofounded by ourselves and others in 1996 and sold in 2004 to ISO Inc. This firm was renamed Verisk Analytics; its health division, Verisk Health, develops and licenses its DCG-HCC based classification system and regression based risk adjustment models under the name DxCG to government and commercial users and to researchers. Although the authors of this paper no longer have direct financial ownership in Verisk Health, they have each consulted and received research funding from Verisk Health, Inc. to refine and update the original DxCG models.

Quantitative comparisons of model performance

For use in Medicare risk adjustment, the DCG-HCC models were chosen over other competing models by CMS researchers in 1996 (Ingber, 1998) and implemented for Part C payment in 2000 and 2004. In head-to-head comparisons of models by independent researchers, such as those conducted by the Society of Actuaries (Dunn et al, 1996; Winkelman and Mehmud 2007), DxCG models consistently perform as well or better than other models (including categorical ones) on standard measures of performance (R^2s, MAPEs and predictive ratios for policy-relevant subgroups – that is, the ability to match average predicted to average actual outcomes in those subgroups). Cid et al (2016) provides a summary of eight different international studies comparing various risk adjustment models, including both categorical and additive models, which confirms the superior predictive power of regression-based additive models.

More generally, DCG-HCC and DxCG models have been repeatedly shown to work well for predicting various outcomes, most prominently cost, but also more clinical outcomes, such as mortality or hospital admissions. These findings hold both when making predictions in general populations and within subpopulations (e.g., among children, for people with diabetes, and in commercial, Medicaid or Medicare populations). (Ash and Ellis, 2012; Cid et al, 2016)

We know of no independent studies that show measureable advantage for any categorical model over the DxCG regression-based models.

Real-world complexity and clinical meaningfulness

The Fuller et al. (2016) paper worries that regression models, especially those with a simple linear structure, will "misprice" a situation in which the presence of two diseases is worse than one might expect from the risk associated with each disease alone. In their example, the presence of A alone leads to costs of $100, and B alone to $300, but a person with both A and B costs $1000 (rather than the $400 you might expect from simple addition). Their concern is unfounded. Regression models easily handle this: A is priced at $100, B at $300, and an "AB

interaction" term is priced at $600, leading to a $1000 (=$100 + $300 +$600) cost estimate for those with both A and B.

Indeed a regression model with each condition category being a distinct clinical entity that contributes to total risk is an elegant framework for describing and studying any imaginable combination of diseases – not just the limited set that a particular group of developers, at some moment in history, felt that they had room to include in their categorical model.

More generally, regression is a powerful, flexible tool, and what Fuller, et al. (2016) calls "categorical models" are, in fact, a class of particularly simple regression models. Consider the current MS-DRGs model with 749 categories (C_1, C_2, … , C_{749}) defined so that every payable hospitalization is classified into exactly one category. Such models are developed (typically, in a large, "benchmark" data set) using a (no-intercept, ordinary least squares) regression as in:

$$PRED = a_1 * C_1 + a_2 * C_2 + \ldots a_{749} * C_{749}$$

The "a" numbers are called coefficients. The regression model will find: a_1 =the average cost of cases in the first category, a_2 =the average cost of cases in the second category, etc. The above equation can be "read" as follows: if a hospitalization's characteristics cause it to be placed (or "grouped") into category i then its expected cost is a_i, the average cost of all cases that grouped into category i in the benchmark data. This is why DRG-type models are also called "groupers."

Thus, the key structural difference between the Fuller et al. (2016) paper's "categorical" and "regression" models is in their "building blocks" (or "atoms"), which both kinds of models principally base on the presence of ICD-9 (or ICD-10) diagnosis codes.

- The "atoms" of a regression model, such as Verisk Health's DxCG, are *individual clinical condition categories, or CCs. Each "case" (here, a person) is described by his or her age, sex and which of 394 distinct CCs are present.* Each diagnosis code maps to a CC. (A small number of codes, such as "ophthalmic complications of diabetes," describe more than one clinical problem and map to more than one CC.) By adding selected interactions as needed, the model can accurately characterize the importance of potentially more than 2^{394} combinations of CCs, more than the number of atoms in the universe!
- The "atoms" of a categorical model, such as the authors' MS-DRGs, are complex, rule-based categories. Each "case" (here a payable hospitalization) maps to exactly one of *749 diagnostic related groups, or DRGs, each of which is a combination of clinical or other patient characteristics*, such as "nervous system neoplasms *with* major complications or comorbidities" or "nervous system neoplasms *without* major complications or comorbidities." Such a model cannot detect differences, for example, in cases of nervous system neoplasms with *different kinds* of serious complications or comorbidities.

Clinical meaningfulness

Regression models are applied using software, just as categorical models are, to predict an outcome for each person. Users are typically not interested in model "mechanics;" they get to see what the model predicts for any individual or group of people. As mentioned above, if predictions for a subgroup of interest do not match actual outcomes, interaction terms can fix this. Indeed, DxCG models already include age-sex interactions with disease (such as, asthma for kids versus adults) and disease-disease interactions (such as, CHF and diabetes), so long as they both make sense to clinicians and actually affect costs (that is, have model coefficients that are statistically distinguishable from zero).

All DCG-HCC models start by organizing diagnosis codes on claims or encounter data into a manageable number of distinct condition categories (CCs). As previously stated, the DxCG software classifies all ICD-9 (and now, ICD-10) codes into 394 condition categories, making it easy to identify, and make predictions for, people with *any combination of clinical characteristics*. Furthermore, if a user of such a model finds that the existing software does not predict well (in a new setting) for *any* subgroup of people, ***S***, that is easily fixed. Specifically, let PRED be the model's original prediction. Any regression software can find the best choices of constants a, b and c to make new predictions:

$$\text{NEWPRED} = a + b*\text{PRED} + c*(\text{flag for people in } \boldsymbol{S}). \qquad (1)$$

Then users can *apply* the new model predictions in new settings, taking advantage of the stability of relationships among all diseases identified in the developer's original benchmark data to make predictions in much smaller data sets. To give a sense of scale: DxCG models are typically developed on data from millions or even tens of millions of people, but the models can be reliably recalibrated as above, so long as there are at least a few thousand people in ***S.***

Indeed, the same idea can be used to make multiple modifications to ensure that a regression model fits *several* new groups of people well (let's call flags for them $\boldsymbol{S}_1$, $\boldsymbol{S}_2$,… , $\boldsymbol{S}_k$,) as in

$$\text{NEWPRED2} = a + b*\text{PRED} + c_1*\boldsymbol{S}_1 + c_2*\boldsymbol{S}_2 + \ldots + c_k*\boldsymbol{S}_k .$$

Further, a regression model can be used to explore differences in costs among people *within* subgroups, for example: the additional implications of renal insufficiency among people with CHF, diabetes and COPD. Categorical models cannot explore differences *within* their pre-specified categories.

In summary, unlike categorical models, regression models predict outcomes both *for* and *within* any subgroup of potential interest to a stakeholder, regardless of whether the original developers included that combination of factors in their model.

Model development

Developing and updating the DCG-HCC models has involved thousands of hours of clinician time over more than two decades; six clinician researchers have coauthored peer-reviewed articles on its development. The rich collaboration between statisticians and clinicians in developing and calibrating early DCG-HCC models is documented in numerous long reports to CMS (then HCFA). (See extensive cites in Ash et al, 1989, Ash et al 2000, Pope et al 2004).

When DCG-HCC models are developed or updated, clinicians and statisticians collaborate to create clinical categories *defined by single illnesses*; they also identify interactions (between diseases or between age or sex and disease) that may be important. The process is iterative – clinicians and modelers (statisticians or econometricians) each suggest and explore issues of potential concern. A typical [hypothetical] dialog might go something like this:

Clinician:	Asthma is different in kids than in adults.
Statistician:	[After running a regression.] Here's what the model tells us about how the expected cost of asthma differs for people under and over age 18.
Clinician:	Well, I wouldn't have guessed those exact numbers, but I expected asthma to be [more/less] expensive for kids, so that seems reasonable.
Statistician:	The difference is pretty solid – the coefficient has a t-statistic of 10.
Conclusion:	OK, we'll add an [Age < 18]*[Asthma] interaction to the model.

The process concludes when the model's clinical clarity and performance characteristics satisfy both groups.

A goal of all risk adjustment models (categorical or regression-based) is to assign an expected outcome to everyone, enabling users to compare an actual outcome, say cost, to what the model predicts, and to make statements like "this group costs 10% more than expected." Because the DxCG model starts by identifying, for each person, exactly which clinical problems are present, its software can readily identify (and price) patients with any combination of distinct clinical problems, with or without age restrictions. With a categorical model it is only easy to see how members in one of its categories are priced (they all get the same price), but not particularly easy to interpret the price for groups that cut across categories.

A final strength of DCG-HCC regression-based models is that they are easy to recalibrate. The basic DxCG modeling structure has for some years relied on 394 distinct clinical categories plus sets of clinical rules embodied in its hierarchies (e.g., a heart attack is more serious than high blood pressure). The model's payment weights undergo major recalibration every few years to recognize shifts in the relationship between medical problems and their costs (such as, the availability of an expensive new drug to care for a previously untreatable problem). Even without recalibration, the predictive power of these models was shown to remain high over a seven year period (Ash and Ellis, 2012). In addition, as described above, it is easy to correct any

mispricing in a new population using a current DxCG model's predictions and simple regression. Since real-world use of any risk adjustment model requires the use of a computer, it is no harder to fit a model such as equation (1) than to compute averages for a categorical model's hundreds of categories.

In conclusion, regression-based models, such as the DxCG hierarchical condition category models, reflect years of collaboration between clinicians and quantitative scientists. They have unsurpassed performance for accurate risk adjustment, are easily understood by stakeholders, including clinicians, and are easy to adapt for new settings.

References

Ash, A. S., Porell, F., Gruenberg, L., Sawitz, E., & Beiser, A. (1989) Adjusting Medicare capitation payments using prior hospitalization data. *Health Care Financing Review*, 10(4), 17-29.

Ash, A. S., Ellis, R.P., Pope, G.C., Ayanian, J.Z., Bates, D.W., Burstin, H., & Yu ,W. (2000) Using diagnoses to describe populations and predict costs. *Health Care Financing Review*, Spring 21(3), 7-28.

Ash, A. S., & Ellis, R. P. (2012) Risk-adjusted payment and performance assessment for primary care. *Medical Care*, 50(8), 643-653.

Cid, C., Ellis, R. P., Vargas, V., Wasem, J & Prieto, L. (2016) Global risk-adjusted payment models. In R. Scheffler, (Ed.) *Handbook of Global Health Economics and Public Policy*. Chapter 11, Vol 1.

Dunn, D. L., Rosenblatt, A., Tiaira, D. A., Latimer, E., Bertko, J., Stoiber, T,…, Busch, S. (1996) *A comparative analysis of methods of health risk assessment*. Final report to the Society of Actuaries.

Ellis, R.P., & Ash, A.S. (1995) Refinements to the Diagnostic Cost Group model. *Inquiry*, Winter, 418-429.

Fuller, R.L., Averill, R.F., Muldoon, J.H. & Hughes, J.S. (2016) Comparison of the properties of regression and categorical risk-adjustment models. *Journal of Ambulatory Care Management*

Ingber, M.J. (1998) The current state of risk adjustment technology for capitation. *Journal of Ambulatory Care Management*, 21(4), 1-28.

Kautter, J., Pope, G.C., Ingber, M., Freeman, S., Patterson, L., Cohen, M., & Keenan, P. (2014), The HHS-HCC risk adjustment model for individual and small group markets under the Affordable Care Act. *Medicare & Medicaid Research Review*, 4(3), E1-E11.

Pope, G.C., Ellis, R.P., Ash, A.S., Liu, C., Ayanian, J.Z., Bates, D.W., … Ingber, M.J., (2000) Principal inpatient diagnostic cost group models for Medicare risk adjustment. *Health Care Financing Review,* Spring 21(3), 93-118.

Van de Ven, W.P.M.M., & Ellis, R.P. (2000) Risk adjustment in competitive health plan markets. In A. Culyer & J. Newhouse, (Eds.) *Handbook in Health Economics* (pp. 755-845), North Holland.

Winkelman, R., & Mehmud, S. (2007) *A comparative analysis of claims-based tools for health risk assessment.* Schaumburg, Ill.: Society of Actuaries.