Statistically Derived Predictive Models

Caveat Emptor

When Flanders and collaborators applied to their own patient population the pneumonia severity index (PSI) that was developed by Fine and colleagues,^{1,2} they discovered discrepancies. The PSI predicted about 2.4 times more deaths than were actually observed.³ This finding is not surprising; it confirms something we already knew. Unlike mathematical representations of basic natural phenomena, such as $E = mc^2$, algorithms derived from statistical modeling reflect only associations embedded within the derivation data set—a snapshot in place and time, not fundamental truth. When transporting statistically derived measures, such as the PSI, to another place and time, users should observe the axiom, “Caveat emptor.”

Fortunately, as Flanders and colleagues knew,³ relatively straightforward methods exist to identify problems such as imperfect calibration. Justice, Covinsky, and Berlin suggest ways to address concerns about what they call the “external validity” of prognostic measures derived from statistical models.⁴ They emphasize two broad concepts: accuracy (whether predictions match outcomes in the original patient sample), which is assessed by examining calibration and discrimination; and generalizability (the extent to which the model provides accurate predictions in different patient samples). Generalizability itself has several components: reproducibility in comparable patient populations and transportability across different dimensions, including place and time.⁴ Justice, Covinsky, and Berlin offer a systematic approach for evaluating the performance of prognostic models, but as they note, the relative importance of different aspects of external validity depends on one’s purpose.⁴

The difficulty arises when the prognostic model’s external validity cannot be determined for its planned use or purpose. In most such instances, the missing ingredients are data. For example, Charleson and collaborators developed their comorbidity index to predict 1-year mortality using information on a cohort of 604 patients admitted to the medical service at New York Hospital during 1 month in 1984.⁵ To validate their comorbidity index, they used information on 685 patients treated between 1962 and 1969 for primary breast cancer at Yale New Haven Hospital. Obviously, 1-year mortality is heavily influenced by which medical therapies are available, and presumably treatments for comorbid conditions had changed during the 15 to 20 years that separate the initial cohort and the validation sample. Exactly why the breast cancer sample was used for validation is unclear, unless it was to provide a severe test of generalizability in quite different patients, but the authors acknowledge that their index required “further evaluation in much larger populations.”⁵ The Charlson Comorbidity Index remains popular today, but its dependence on 1984 therapeutic outcomes must be considered when applying it to 1999 data.

An example relevant to health policy is the diagnosis-related group (DRG) classification system, which since late 1983 has determined Medicare’s prospective payment for most hospitalized patients.⁶ Periodically, DRG payment weights must be calibrated to hospital costs nationwide. Given the importance of setting accurate payments, this calibration requires timely data.

Although its basic diagnostic groupings were first derived clinically, the final DRG classification scheme resulted from statistical modeling of a database containing more than 1.5 million Medicare discharges.⁷ Annually, for the October 1 start of the federal fiscal year, DRG payment weights and sometimes the fundamental structure of the DRGs themselves are recalibrated using the most recent available Medicare data. Sometimes, major changes are made. For example, in the initial version, some DRGs were created by splitting certain patient groups into those younger than and those older than 69 years; that split was soon abandoned when analyses found that the age break had little statistical value. Other major alterations in the classification system include the creation of new categories for continuous mechanical ventilation and for operating room procedures unrelated to the principal diagnosis. These changes, although justified statistically, have complicated DRG-based comparisons over time.

The Medicare data available for this annual recalibration are inevitably 1 or 2 years old. The data, therefore, cannot capture the cost implications of rapidly disseminated, new, diagnostic or therapeutic technologies, such as thrombolytic therapy for acute myocardial infarction, endoscopic surgery, or novel transplant techniques. To some extent, DRG payment weights are always playing catch-up with the way hospital-based medicine is practiced.

Another Medicare situation that involves the use of outdated or inadequate data for calibrating prognostic payment models could have significant consequences. The 1997 Balanced Budget Act (P.L. 105-33) mandated substantial changes in Medicare, including the expansion of capitated care. To lessen any financial incentives that would discourage the enrollment of sick and disabled people, the Balanced Budget Act requires Medicare to link capitation payments to health status, or to “risk adjust” the reimbursement, by January 1, 2000.⁸⁻¹⁰ In this context, to risk adjust means to adjust payment so it reflects the expected cost of care based on the patient’s clinical characteristics, for example, to pay more for a lung cancer patient than for a person without known illness. The major problem is that the data for setting and calibrating these payments are inadequate.

Historically, Medicare did not require capitated plans to submit records on hospitalizations or outpatient services, to minimize the administrative burden. The Balanced Budget Act changed that practice by allowing Medicare to require that managed care organizations report diagnosis

and procedure codes for individual patients. To meet the start date of January 1, 2000, however, the only diagnostic data available for calibrating payment levels come from bills submitted through fee-for-service Medicare. Because of this and other data concerns,^{9,10} Medicare currently plans to risk adjust year 2000 capitated payments using statistical models derived from 1995 and 1996 data from hospitalized, fee-for-service patients.⁹ Thus, Medicare's risk-adjusted, capitated payments will reflect 5-year-old data from a care system with a fundamentally different philosophy than its intended target—managed care.

Fixing problems in statistical models that are caused by the available data can be relatively straightforward. For example, Flanders and colleagues fixed the PSI so it could be used for their purpose with a simple statistical patch—they recalibrated the PSI using logistic regression. Of course, their recalibrated PSI now is inextricably linked to the content and setting of *its own* data. Each recasting of a prognostic model is thus essentially unique. Just as the annual updating of the DRGs and their payment weights has created more than 15 versions, the continual recalibrating of prognostic models like the PSI could produce a bewildering array of methods. Subtle variations in the models could hamper efforts of investigators to replicate or compare findings.

In 1991, concerned about the proliferation of "black box" medical information systems (methods whose inner workings were kept secret, typically by their proprietary developers), I recommended instituting a formal, national process to evaluate these measures.¹¹ Since payments and reports on provider quality are increasingly based on such statistically derived methods, a national clearinghouse that could rate these systems objectively—for example, by examining their reliability, accuracy, and external validity—seemed warranted. A clearinghouse could also maintain catalogs of different versions of the models, delineating their pedigrees and data specifications.

A national clearinghouse is, however, even more unlikely to be established now than it was in the early 1990s. Health plans and large chains of providers increasingly employ their own, homegrown methods to profile their performance, and many such organizations refuse to share details of their methods. Nevertheless, the research community itself should adopt guidelines for published reports employing statistically derived models.

First, all publications should clearly state the version of the model employed. For example, many authors of stud-

ies using DRGs to control for hospital case mix do not indicate the version of the DRGs, which could be done by stating the year the version was adopted by the Health Care Financing Administration. Second, when appropriate, as in the work by Flanders and colleagues,³ investigators should examine the discrimination and calibration of the model in their own data set and report those results. If investigators modify these models in any way, for example, by recalibrating them, they should describe all the changes. Finally, investigators who develop statistically derived models should report explicit details about the derivation database and population sample, including any speculation about unique features that could compromise comparisons with other samples. Though buyers of statistically derived models must still beware, these strategies should at least produce a more informed consumer.—**LISA I. IEZZONI, MD, MSc**, *Division of General Medicine and Primary Care, Department of Medicine, Harvard Medical School, Boston, Mass.*

REFERENCES

1. Fine EM, Hanusa BH, Lave JR, et al. Comparison of a disease-specific and a generic severity of illness measure for patients with community acquired pneumonia. *J Gen Intern Med.* 1995;10:359-68.
2. Fine EM, Singer DE, Hanusa BH, Lave JR, Kapoor WN. Validation of a pneumonia prognostic index using the MedisGroups comparative hospital database. *Am J Med.* 1993;94:153-9.
3. Flanders WD, Tucker G, Krishnadasan A, et al. Validation of the pneumonia severity index (PSI): importance of study specific recalibration. *J Gen Intern Med.* 1999;14:333-40.
4. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med.* 1999;130:515-24.
5. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *J Chron Dis.* 1987;40:373-83.
6. Vladeck BC. Medicare hospital payment by diagnosis-related groups. *Ann Intern Med.* 1984;100:576-91.
7. Fetter RB, Shin Y, Freeman JH, Averill R, Thompson J. Case mix definition by diagnosis related groups. *Med Care.* 1980;18(suppl):1-53.
8. Department of Health and Human Services, Health Care Financing Administration. Medicare program; establishment of the Medicare+Choice program; final rule. *Fed Reg.* 1998;63(123):35004-7.
9. Department of Health and Human Services, Health Care Financing Administration. Medicare program; request for public comments on implementation of risk adjusted payment for the Medicare+Choice program and announcement of public meeting. *Fed Reg.* 1998; 63(173):47506-13.
10. Iezzoni LI, Ayanian JZ, Bates DW, Burstin H. Paying more fairly for Medicare capitated care. *N Engl J Med.* 1998;339:1933-8.
11. Iezzoni LI. "Black box" medical information systems: a technology needing assessment. *JAMA.* 1991;265:3006-7.