American Journal of Epidemiology

Vol. 154, No. 9
Printed in U.S.A.

Performance of Comorbidity Scores to Control for Confounding in Epidemiologic Studies using Claims Data

Sebastian Schneeweiss,[1,2] John D. Seeger,[2] Malcolm Maclure,[2,3] Philip S. Wang,[1,2] Jerry Avorn,[1] and Robert J. Glynn[1,4]

Comorbidity is an important confounder in epidemiologic studies. The authors compared the predictive performance of comorbidity scores for use in epidemiologic research with administrative databases. Study participants were British Columbia, Canada, residents aged ≥65 years who received angiotensin-converting enzyme inhibitors or calcium channel blockers at least once during the observation period. Six scores were computed for all 141,161 participants during the baseline year (1995–1996). Endpoints were death and health care utilization during a 12-month follow-up (1996–1997). Performance was measured by using the *c* statistic ranging from 0.5 for chance prediction of outcome to 1.0 for perfect prediction. In logistic regression models controlling for age and gender, four scores based on the *International Classification of Diseases*, Ninth Revision (ICD-9) generally performed better at predicting 1-year mortality ($c = 0.771$, $c = 0.768$, $c = 0.745$, $c = 0.745$) than medication-based Chronic Disease Score (CDS)-1 and CDS-2 ($c = 0.738$, $c = 0.718$). Number of distinct medications used was the best predictor of future physician visits ($R^2 = 0.121$) and expenditures ($R^2 = 0.128$) and a good predictor of mortality ($c = 0.745$). Combining ICD-9 and medication-based scores improved the *c* statistics (1.7% and 6.2%, respectively) for predicting mortality. Generalizability of results may be limited to an elderly, predominantly White population with equal access to state-funded health care. *Am J Epidemiol* 2001;154:854–64.

comorbidity; confounding factors (epidemiology); databases; epidemiologic studies; health services

Comorbidity scores can be useful tools for controlling for confounding in epidemiologic analyses in which claims-based data are used. However, little is known about the relative performance of various available comorbidity scores in predicting a variety of outcomes (1). Particular measures often seem to be chosen for convenience rather than performance. The construct "comorbidity" reflects the aggregate effect of all clinical conditions a patient might have, excluding the disease of primary interest (2). Because there is no "gold standard," researchers validate measures of comorbidity by how well they predict worse health outcomes, more health care utilization, and increased health care expenditures.

Received for publication October 20, 2000, and accepted for publication June 26, 2001.

Abbreviations: CDS, Chronic Disease Score; ICD-9, *International Classification of Diseases,* Ninth Revision; OR, odds ratio; RR, relative risk; SD, standard deviation.

[1] Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA.
[2] Department of Epidemiology, Harvard School of Public Health, Boston, MA.
[3] Pharmacare, Ministry of Health, British Columbia, Canada.
[4] Department of Preventive Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA.
Reprint requests to Dr. Sebastian Schneeweiss, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, 221 Longwood Avenue (BLI-341), Boston, MA 02115 (e-mail: schneeweiss@post.harvard.edu).

The predictive performance of claims-based comorbidity scores depends on several factors, including 1) the clinical conditions included in a score and their relative weights; 2) the distribution of comorbid conditions in the source population; 3) the endpoint of a study, for example, 1-year mortality; and 4) the accuracy of the administrative data (3). The predictive performance of two scores can validly be compared when factors 2–4 are held constant. Several studies have explored the predictive validity of comorbidity measures in claims data (4–13). However, only a few publications compared the performance of two comorbidity scores in the same populations and for the same endpoints (11, 12, 14). We are unaware of any direct comparison of medication-based versus diagnosis-based scores or more than two scores in the same population.

In this study, we compared the performance of six claims-based comorbidity scores in predicting 1-year mortality, long-term-care admission, number of hospitalizations, physician visits, and expenditures for physician services. The study population was a cohort of British Columbia, Canada, residents aged 65 years or more who had hypertension.

MATERIALS AND METHODS

The study population included all British Columbia residents aged 65 years or more on March 31, 1996, for whom at least one health care encounter was paid for by the

Ministry of Health (for prescription medication, medical service, or hospitalization) during the 4 months prior to the baseline year (April 1, 1995–March 31, 1996). As part of a larger policy study (15), data on all filled prescriptions, health care utilization, and expenditures were available for all patients who had filled at least one prescription for an angiotensin-converting enzyme inhibitor or calcium channel blocker from January 1, 1995, to December 31, 1997. Patients who died or were admitted to long-term care during the baseline period were excluded. The cohort of eligible patients (n = 141,161) was followed for 1 year after baseline (April 1, 1996–March 31, 1997). Comorbidity was assessed during the baseline year, and all endpoints were assessed during the follow-up year.

Scores

Original research on the metric properties of comorbidity indices for claims data was identified by a literature search using MEDLINE (National Library of Medicine, Bethesda, Maryland) and HealthStar (HealthStar, Inc., Long Beach, California) databases, bibliographies, and expert consultations. We identified six distinct indices of comorbidity for use in administrative databases (4, 7–9, 11, 14). Four of the six scores use diagnostic information from *International Classification of Diseases*, Ninth Revision (ICD-9) codes and are based on the Charlson index originally designed for clinical data (16). Two of the scores are based on outpatient drug utilization data.

Diagnosis-based scores. The Charlson index is a list of 19 conditions; each is assigned a weight (1 to 6). The Charlson index score is the sum of the weights for all conditions that a patient has. Although the index might seem rather simple, it was associated with a 2.3-fold (95 percent confidence interval: 1.9, 2.8) increase in the 10-year risk of death per increment in comorbidity level in a cohort of 685 breast cancer patients (16), and similar results were found for postoperative survival in patients with hypertension or diabetes (17).

For the Deyo and Romano implementations of the Charlson index, we used the corresponding sets of five-digit ICD-9-CM (Clinical Modification) diagnoses, as delineated in these authors' original publications (5, 8). These two scores differ only modestly in the ICD-9-CM codes that map the Charlson index conditions (5).

For the D'Hoore implementation of the Charlson comorbidity index, we used the first three digits of the ICD-9 code, as described by D'Hoore et al. (9). The Ghali adaptation of the Charlson index was calculated with the reduced set of diagnoses specified by Ghali et al. (11).

The four scores were calculated by using ICD-9 codes derived from all hospital discharges, which can contain up to 16 diagnoses. In addition to these original scores based on hospitalization only, we also calculated scores based on the diagnoses associated with all inpatient and outpatient physician services or procedures received during the baseline year.

The original Charlson weights were applied to the Deyo, Romano, and D'Hoore scores. The published weights were applied to the Ghali score (refer to table 4 in reference (1)).

As a simple measure, we also used the number of distinct diagnoses, that is, different first-three-digit ICD-9 codes during the baseline year. There were two categories, hospital discharge diagnoses and hospital plus ambulatory diagnoses.

Prescription-medication-based scores. For the Chronic Disease Score (CDC), outpatient pharmacy dispensing data are used to assign patients to chronic disease groups. An integer weight is given to each comorbidity category represented by selected medication classes, and all weights are summed to obtain an overall score. The CDS was developed by an interdisciplinary expert group of researchers and practitioners and was refined after several pilot studies. CDS-1 was tested among 122,911 Group Health Cooperative (Washington State) enrollees. A multivariate logistic regression model showed that with an increasing CDS-1 score, the probabilities of 1-year hospitalization and of 1-year mortality increased steadily. Compared with patients who were in the lowest CDS-1 score category, those in the highest category (7+) had a 10-fold higher probability of dying in the next year. An extended version of the score, CDS-2 (14), was designed specifically to predict future health care utilization.

To calculate the CDS-1 score, we followed the original coding (7). For the CDS-2 score, the published weights used to predict primary care visits were adopted. (14). Drugs that have become available since 1992 were assigned to an appropriate category based on the condition for which the medication is prescribed. For example, only cimetidine was originally specified as an indicator for ulcer disease, and we expanded this list to include any H (histamine)$_2$ antagonist or proton pump inhibitor. For drugs that were available when the score was developed but for which their indications have since been expanded to include one of the scored chronic diseases, the disease categories were not changed for that drug (e.g., methotrexate for cancer but now used more frequently for rheumatoid arthritis).

We used number of distinct prescription drugs (distinct chemical entities) dispensed during the baseline year as a crude comorbidity measure. Medications whose first eight digits of the American Hospital Formulary Services code (18) were equal were considered the same substance.

Other utilization measures. Two other simple utilization measures were also considered as predictors: 1) Number of hospitalizations for any reason and any length during the baseline year. Elective hospitalizations and unplanned emergency hospitalizations were differentiated. 2) Number of physician visits for any reason during the baseline year.

Endpoints

The primary endpoint was mortality during the follow-up year. Secondary endpoints were long-term-care admissions, hospitalizations (elective and emergency), number of physician visits (including services in hospitals), and expenditures for physician services during the follow-up year. Expenditures were measured by payments by the provincial government. For patients who left the cohort for reasons other than dying during the follow-up year, numbers of physician visits and expenditures were extrapolated to an annual count (19). The rate of emigration from British Columbia is very low among residents aged 65 years or more (20).

Data quality

In British Columbia, pharmacists enter pharmacy dispensing data—including medication, strength, and number of units—into a computer network when a prescription is filled, and underreporting and misclassification appear to be minimal (21). Although previous reports indicate reasonable levels of accuracy and completeness of diagnostic coding (22), misclassification of ICD-9 diagnoses is probably similar to that found in research in which other administrative databases are used (23–26). British Columbia pays all medication and medical services costs for residents aged 65 years or more. Data on medical services include accurate information on the amount paid in Canadian dollars.

Data analysis

For each endpoint, three baseline regression models were fitted to the data by modeling endpoints as a function of age, gender, and age plus gender combined. For each of the six comorbidity scores, models were constructed containing only the score as well as the score plus age and gender. Dichotomous endpoints (mortality, long-term-care admissions) were modeled by fitting logistic regression models; *c* statistics (i.e., the area under the receiver operating characteristic (ROC) curve) were calculated as measures of discrimination (27). The *c* statistic ranges from 0 to 1, with 1 indicating a perfect prediction and 0.5 a chance prediction; for example, the Framingham Heart Study could predict the incidence of coronary heart disease based on age, blood pressure, smoking, diabetes, and low density and high density lipoprotein cholesterol levels with a *c* statistic of 0.77 (28). It has been suggested that *c* statistics of 0.7–0.8 could be considered acceptable and those of 0.8–0.9 excellent (29); higher values are rarely observed and are described as outstanding. Asymptotic 95 percent confidence limits were reported for *c* statistics (30). Because multiple hospitalizations occurred in less than 5 percent of patients during follow-up, we categorized patients as those without and those with one or more hospitalizations. For continuous outcomes (expenditures for physician services), we fitted linear regression models and reported R^2 statistics to reflect the proportions of explained variance (31). Since number of physician visits per year varied widely around a mean of 10.9 (standard deviation (SD), 12.4), we considered it a continuous variable. Expenditure and visit data were considerably skewed to the right and therefore were log-transformed (32). Predictive performance should not be compared across outcomes but across scores within outcomes. Spearman's correlation coefficients with two-sided *p* values were calculated among scores and utilization measures during the baseline year.

Another way to quantify the performance of scores is to estimate how much confounding by comorbidity would be avoided by adjusting for each of the six scores, assuming an underlying null association between an exposure and outcome. Since the scores represent measurable confounding by comorbidity, it can be controlled for in stratified analyses. The true amount of confounding caused by comorbidity might be larger but remains unknown. The difference in confounding that can be adjusted between scores reflects the relative capacity of each score to adjust for confounding. Scores that perform equally may do so by controlling for different qualities of comorbidity; that is, the scores are not necessarily nested within each other.

To determine how much confounding would be avoided by adjusting for each score, we used actual outcome data and the observed associations between scores and outcome, and we considered various assumptions about the prevalence of exposure and the exposure-comorbidity association. For simplicity, we assumed a dichotomous exposure and a dichotomous comorbidity measure. The apparent or crude relative risk (RR) of an exposure (E)-outcome (O) association in the presence of confounding (crude RR_{EO}) is related to the associations between confounder (C) and exposure (OR_{EC}) as well as confounder and outcome (RR_{CO}; refer to the Appendix). To ensure comparability, and on the basis of the observed distribution of scores, we dichotomized all six scores by choosing cutpoints closest to the 75th percentile. That is, only 25 percent of patients with the highest scores were coded as having a notable degree of comorbidity. We used the observed prevalence of comorbidity Pr(*C*) and the observed confounder-mortality association and varied OR_{EC} from 0.2 to 8. The prevalence of exposure Pr(*E*) was varied between 0.1 and 0.3. The underlying exposure-outcome association was assumed to be constant, with $RR_{EO} = 1$.

RESULTS

Population

At the beginning of the baseline year, the population of 141,161 patients was on average aged 75.4 years (SD, 6.7), and 58 percent were female. The distributions of the comorbidity indices during the baseline period are shown in table 1. During the follow-up year, the average numbers of elective hospitalizations (0.1; SD, 0.3) and emergency hospitalizations (0.2; SD, 0.4) were unchanged, and the number of physician visits increased slightly (10.9; SD, 12). A total of 1,221 Can $ was spent on average per patient per year (SD, 1,627). During the follow-up year, 5,569 deaths occurred, and 3,317 patients were admitted to long-term-care facilities. No study patients migrated permanently out of the province.

Correlations at baseline

Correlations between ICD-9-based and medication-based scores were 0.31 or lower (table 2). Medication-based scores were highly correlated with number of distinct medications received ($r > 0.6$) and weakly correlated with number of emergency hospitalizations during baseline ($r < 0.20$). Conversely, ICD-9-based scores were more highly correlated with hospitalizations ($r \geq 0.30$) but not as well correlated with number of medications ($r \leq 0.35$). Number of physician visits correlated highly with number of different ICD-9 diagnoses ($r = 0.76$ for hospital and ambulatory codes combined).

Performance

The Romano adaptation of the Charlson index performed best in predicting 1-year mortality; the *c* statistic was 0.771 in

TABLE 1. Distributions of six comorbidity scores and several utilization measures during the baseline year, British Columbia, Canada, April 1995–March 1996

Score/measure (reference no.)	Mean (standard deviation)	% with 0	Median	75th percentile	Maximum
CDS*-1 (7)	4.4 (2.3)	6.4	4.0	6	19.0
CDS-2 (14)	3.1 (1.3)	0.0	2.8	3.6	12.1
Deyo (8)	0.5 (1.1)	73.8	0.0	1	12.0
D'Hoore (9, 10)	1.2 (1.8)	56.1	0.0	2	18.0
Ghali (11)	0.4 (1.2)	86.7	0.0	0	11.0
Romano (5, 6)	0.5 (1.1)	73.3	0.0	1	14.0
No. of nonemergency hospitalizations	0.1 (0.5)	88.6	0.0	0	39.0
No. of emergency hospitalizations	0.2 (0.6)	85.2	0.0	0	15.0
No. of distinct prescription drugs†	7.4 (5.1)	1.7	6.0	10	55.0
No. of distinct ICD-9 diagnoses‡	3.5 (2.7)	6.1	3.0	5	39.0
No. of physician visits	8.8 (8.6)	5.3	6.0	12	184.0

* CDS, Chronic Disease Score.
† Prescription medications that have different chemical structures but may be part of the same therapeutic group.
‡ *International Classification of Diseases,* Ninth Revision (ICD-9) diagnoses whose first three digits differ.

a model including age and gender (table 3). This finding represents an improvement of 0.09 (11.7 percent) over an age-and-gender model alone ($c = 0.681$). Deyo's version performed similarly, but the three-digit ICD-9-based D'Hoore score and Ghali's adaptation seemed to perform less well. Both CDS-1 and CDS-2 did not perform as well (table 3).

Performance for predicting long-term-care admissions was generally better, and the rank order was the same (Romano (5, 6) > Deyo (8) > D'Hoore (9, 10) > Ghali (11) > CDS-1 (7) > CDS-2 (14)); however, age contributed considerably to the prediction (table 4). Future emergency hospital admissions were best predicted by number of medications prescribed during the baseline year compared with ICD-9-based scores, followed by medication-based scores. Number of distinct medications seemed to be the best predictor for future physician visits and expenditures for physician ser-

TABLE 2. Spearman's correlation coefficients of six comorbidity scores and selected utilization measures*,† during the baseline year, British Columbia, Canada, April 1995–March 1996

	CDS‡-1 (7)	CDS-2 (14)	Deyo (8)	D'Hoore (9, 10)	Romano (5, 6)	Ghali (11)	No. of prescription drugs§	No. of diagnoses¶	No. of non-emergency hospital-izations	No. of emergency hospital-izations
CDS-2	0.653									
Deyo	0.296	0.293								
D'Hoore	0.298	0.306	0.587							
Romano	0.305	0.301	0.892	0.594						
Ghali	0.240	0.202	0.659	0.409	0.654					
No. of prescription drugs‡	0.646	0.779	0.343	0.349	0.351	0.275				
No. of diagnoses§	0.257	0.327	0.287	0.319	0.289	0.251	0.422			
No. of elective hospitalizations	0.135	0.170	0.333	0.231	0.348	0.263	0.218	0.113		
No. of emergency hospitalizations	0.219	0.236	0.470	0.321	0.477	0.490	0.308	0.218	0.222	
No. of physician visits	0.298	0.371	0.321	0.341	0.332	0.273	0.470	0.711	0.288	0.313

* All p values of the reported Spearman's correlation coefficients, <0.0001.
† Number(s) in parentheses, reference number(s).
‡ CDS, Chronic Disease Score.
§ Prescription medications that have different chemical structures but may be part of the same therapeutic group.
¶ *International Classification of Diseases*, Ninth Revision diagnoses whose first three digits differ.

TABLE 3. Prediction of 1-year mortality by six comorbidity scores* measured 1 year earlier, British Columbia, Canada, 1995–1997

Score (reference no.)	Model†	Continuous score		Binary score‡		Difference
		c statistic	95% confidence interval	*c* statistic	95% confidence interval	
CDS§-1 (7)	Age + gender + CDS-1	0.738	0.731, 0.744	0.721	0.714, 0.728	0.017
CDS-2 (14)	Age + gender + CDS-2	0.718	0.711, 0.725	0.715	0.708, 0.722	0.003
Deyo (8)	Age + gender + Deyo	0.768	0.762, 0.775	0.757	0.751, 0.763	0.011
D'Hoore (9, 10)	Age + gender + D'Hoore	0.745	0.739, 0.752	0.719	0.712, 0.726	0.027
Romano (5, 6)	Age + gender + Romano	0.771	0.764, 0.777	0.758	0.751, 0.764	0.013
Ghali (11)	Age + gender + Ghali	0.745	0.738, 0.752	0.742	0.735, 0.749	0.003

* Scores modeled as continuous variables and as binary variables.
† Age was entered as a linear term into all models.
‡ The cutpoint for the dichotomous transformation of all six scores was chosen as the one closest to the 75th percentile of each score.
§ CDS, Chronic Disease Score.

vices but also a good predictor for long-term-care admissions. CDS-2 performed poorly in predicting physician visits, despite the fact that its weights were specifically designed to perform well for this endpoint.

Performance of scores based on hospital and ambulatory ICD-9 codes was only slightly better than using hospital discharge codes alone (table 4). We observed a 1.3 percent improvement in the Romano score based on hospital discharge diagnoses ($c = 0.757$) when compared with the Romano score based on both ambulatory and hospital data ($c = 0.770$). Only number of distinct diagnoses performed better when hospital discharge diagnoses, and not ambulatory codes, were used.

In the regression analyses, each score and age were modeled as linear terms. When age and the scores were divided into tertiles and were included in the models as ordinal variables, their predictive performance for mortality decreased marginally (<0.5 percent). Scores were also divided into two categories, with cutpoints chosen to be closest to the 75th percentile. Doing so decreased performance an average of 1.7 percent except for the D'Hoore score, which decreased by 2.7 percent (table 3). When quadratic terms of the scores were added to regression models, the *c* statistics improved less than 0.5 percent for all scores except those for the CDS-2, which improved by 0.9 percent.

Because ICD-9-based and medication-based scores were not strongly correlated, we fitted models including both types of scores to improve the predictive value (table 5). Combining the CDS-1 with ICD-9-based scores improved the prediction for all outcomes. The improvements in *c* statistics were smaller for ICD-9-based scores (e.g., Deyo + CDS-1 = 2 percent; Romano + CDS-1 = 1.7 percent at predicting mortality) but larger for medication-based scores (e.g., CDS-1 + Romano = a 6.2 percent improvement over CDS-1 alone). The combination of ICD-9-based scores and number of medications performed equally well or better than the combination of ICD-9-based scores and CDS-1 score (table 5).

Figure 1 shows the percentage of confounding bias that would be controlled by each of the comorbidity scores, assuming there is no underlying association between an exposure and outcome ($RR_{EO} = 1$). There is no confounding bias if there is no association between exposure and comorbidity ($OR_{EC} = 1$). If the exposure is associated with comorbidity and the odds ratio is 3.0, then the comorbidity, as measured by the Romano or Deyo score, would cause a bias of 47 percent, and the crude exposure-outcome relative risk would be estimated to be 1.47. Because this represents the amount of measured confounding, we assume that this confounding can be adjusted by using the Romano or Deyo score in a stratified analysis. On the other hand, the medication-based CDS-1 score would control only 28 percent of the bias, or 60 percent of that controlled by the Romano or Deyo score. For rare exposures ($Pr(E) = 0.1$), the Romano and Deyo scores adjusted equally well and outperformed the medication-based scores by about 40 percent. D'Hoore's score performed only slightly better than the medication-based scores. For more frequent exposures ($Pr(E) = 0.3$), the relative order was unchanged.

Because of the extremely skewed distribution of the Ghali score, a binary cutpoint occurred between a raw score of 0 and 1 and thus led to a prevalence of confounding of only 13 percent compared with 25 percent for all other scores. Therefore, to avoid unfair comparisons, figure 1 does not show the Ghali score.

DISCUSSION

In an elderly population, ICD-9-based comorbidity scores tended to perform better than medication-based scores in predicting future mortality and morbidity. This finding is consistent with the hypothesis that diagnosed conditions not treated by drugs, and pairs of diagnosed conditions treated by only one drug (e.g., hypertension and angina treated by one calcium channel blocker), are important to count. The

TABLE 4. Prediction of 1-year mortality and 1-year health care utilization by six comorbidity scores and several measures of utilization measured 1 year earlier, British Columbia, Canada, 1995–1997

Score/measure (reference no.) and model*	Binary outcome (*c* statistic)				Continuous outcome (R^2)	
	Mortality	Nonemergency hospitalization	Emergency hospitalization	Long-term-care admissions	Physician visits	Expenditures for physician services
Demographics						
Age	0.667	0.527	0.601	0.776	0.007	0.005
Gender	0.543	0.528	0.515	0.553	0.000	0.000
Age + gender	0.681	0.544	0.605	0.776	0.007	0.006
CDS†-1 (7)						
CDS-1	0.659	0.561	0.590	0.597	0.049	0.048
Age + gender + CDS-1	0.733	0.575	0.637	0.792	0.055	0.053
CDS-2 (14)						
CDS-2	0.633	0.579	0.605	0.601	0.060	0.064
Age + gender + CDS-2	0.718	0.588	0.645	0.787	0.067	0.070
Deyo (8)						
Deyo	0.694 (0.656)‡	0.580 (0.562)	0.601 (0.581)	0.644 (0.639)	0.059 (0.045)	0.050 (0.032)
Age + gender + Deyo	0.768 (0.757)	0.598 (0.589)	0.653 (0.649)	0.812 (0.815)	0.064 (0.051)	0.054 (0.037)
D'Hoore (9, 10)						
D'Hoore	0.675 (0.651)	0.578 (0.563)	0.597 (0.578)	0.669 (0.635)	0.073 (0.043)	0.063 (0.031)
Age + gender + D'Hoore	0.745 (0.752)	0.589 (0.590)	0.639 (0.645)	0.806 (0.809)	0.076 (0.049)	0.066 (0.036)
Romano (5, 6)						
Romano	0.696 (0.657)	0.585 (0.563)	0.604 (0.582)	0.649 (0.641)	0.062 (0.046)	0.052 (0.033)
Age + gender + Romano	0.771 (0.757)	0.603 (0.591)	0.655 (0.649)	0.813 (0.816)	0.067 (0.051)	0.056 (0.037)
Ghali (11)						
Ghali	0.649 (0.618)	0.552 (0.540)	0.577 (0.560)	0.622 (0.603)	0.042 (0.033)	0.031 (0.022)
Age + gender + Ghali	0.745 (0.733)	0.576 (0.570)	0.642 (0.636)	0.796 (0.796)	0.046 (0.037)	0.034 (0.026)
No. of distinct prescription drugs§						
No. of prescription drugs	0.677	0.598	0.632	0.634	0.118	0.124
Age + gender + no. of prescription drugs	0.745	0.609	0.668	0.798	0.121	0.128
No. of distinct ICD-9 diagnoses¶						
No. of diagnoses	0.626 (0.659)	0.545 (0.585)	0.555 (0.602)	0.608 (0.676)	0.051 (0.057)	0.036 (0.042)
Age + gender + no. of diagnoses	0.721 (0.748)	0.564 (0.605)	0.619 (0.657)	0.794 (0.822)	0.056 (0.061)	0.041 (0.046)
No. of elective hospitalizations						
No. of hospitalizations	0.562	0.557	0.533	0.574	0.021	0.017
Age + gender + no. of hospitalizations	0.704	0.589	0.619	0.794	0.028	0.023
No. of emergency hospitalizations						
No. of emergency hospitalizations	0.634	0.555	0.593	0.651	0.044	0.033
Age + gender + no. of emergency hospitalizations	0.732	0.583	0.658	0.808	0.049	0.037
No. of physician visits						
No. of physician visits	0.627	0.527	0.533	0.567	0.055	0.03
Age + gender + no. of physician visits	0.727	0.550	0.612	0.786	0.061	0.035

* All continuous variables, including age and comorbidity scores, were entered as linear terms into the models.
† CDS, Chronic Disease Score.
‡ Numbers in parentheses, corresponding statistics when only *International Classification of Diseases*, Ninth Revision (ICD-9) codes from hospital discharge diagnoses were used.
§ Prescription medications that have different chemical structures but may be part of the same therapeutic group.
¶ ICD-9 diagnoses whose first three digits differ based on ambulatory and hospital diagnoses; values in parentheses based on hospital discharge diagnoses only.

contrary hypothesis, that medication-based scores would capture important diagnoses that failed to be coded into the database, was not supported. It has been shown that the more ill patients are, the less likely that some comorbid conditions will be treated (33, 34). In particular, medications that have some preventive effects (e.g., oral antidiabetics or

TABLE 5. Prediction of 1-year mortality and 1-year health care utilization by models in which ICD-9*-based and drug-based comorbidity scores measured 1 year earlier were combined, British Columbia, Canada, 1995–1997

Model (reference no.)†	Binary outcome (*c* statistic)				Continuous outcome (R^2)	
	Mortality	Nonemergency hospitalization	Emergency hospitalization	Long-term-care admissions	Physician visits	Expenditures for physician services
Age	0.667	0.527	0.601	0.776	0.007	0.005
Gender	0.543	0.528	0.515	0.553	0.000	0.000
Age + gender	0.681	0.544	0.605	0.776	0.007	0.006
Age + gender + CDS*-1 + Deyo (8)	0.782	0.605	0.661	0.815	0.088	0.079
Age + gender + CDS-1 + D'Hoore (9, 10)	0.766	0.597	0.651	0.811	0.098	0.089
Age + gender + CDS-1 + Romano (5, 6)	0.783	0.608	0.662	0.817	0.090	0.080
Age + gender + CDS-1 + Ghali (11)	0.768	0.590	0.655	0.803	0.076	0.067
Age + gender + CDS-1 + no. of elective hospitalizations	0.747	0.604	0.643	0.803	0.068	0.063
Age + gender + CDS-1 + no. of emergency hospitalizations	0.760	0.595	0.668	0.814	0.080	0.070
Age + gender + CDS-1 + no. of diagnoses‡	0.751	0.581	0.640	0.802	0.088	0.075
Age + gender + CDS-1 + no. of physician visits	0.753	0.576	0.638	0.796	0.086	0.066
Age + gender + no. of prescription drugs§ + Deyo (8)	0.781	0.624	0.680	0.818	0.137	0.137
Age + gender + no. of prescription drugs + D'Hoore (9, 10)	0.767	0.618	0.674	0.814	0.145	0.145
Age + gender + no. of prescription drugs + Romano (5, 6)	0.783	0.627	0.680	0.819	0.138	0.138
Age + gender + no. of prescription drugs + Ghali (11)	0.770	0.616	0.678	0.807	0.131	0.132
Age + gender + no. of prescription drugs + elective hospitalizations	0.751	0.626	0.670	0.808	0.126	0.130
Age + gender + no. of prescription drugs + emergency hospitalizations	0.759	0.617	0.684	0.815	0.129	0.131
Age + gender + no. of prescription drugs + no. of diagnoses	0.751	0.609	0.668	0.804	0.133	0.132
Age + gender + no. of prescription drugs + no. of physician visits	0.752	0.611	0.670	0.800	0.130	0.128

* ICD-9, *International Classification of Diseases*, Ninth Revision; CDS, Chronic Disease Score; for information about CDS-1, see reference (7).

† All continuous variables, including age and comorbidity scores, were entered as linear terms into the models.

‡ ICD-9 diagnoses whose first three digits differ based on ambulatory and hospital diagnoses; values in parentheses based on hospital discharge diagnoses only.

§ Prescription medications that have different chemical structures but may be part of the same therapeutic group.

lipid-lowering drugs) are prescribed less frequently for very ill patients, causing them to seem healthier according to their medication-based scores.

The enhanced Chronic Disease Score (CDS-2), which was designed to predict future physician visits, performed better than its predecessor (CDS-1) in predicting visits and expenditures. However, both were outperformed by number of distinct medications received during the baseline year, which was the best predictor of future physician services and expenditures, and it performed better than both CDSs in predicting mortality, hospitalizations, and long-term-care admissions, perhaps because conversion of number of dis-

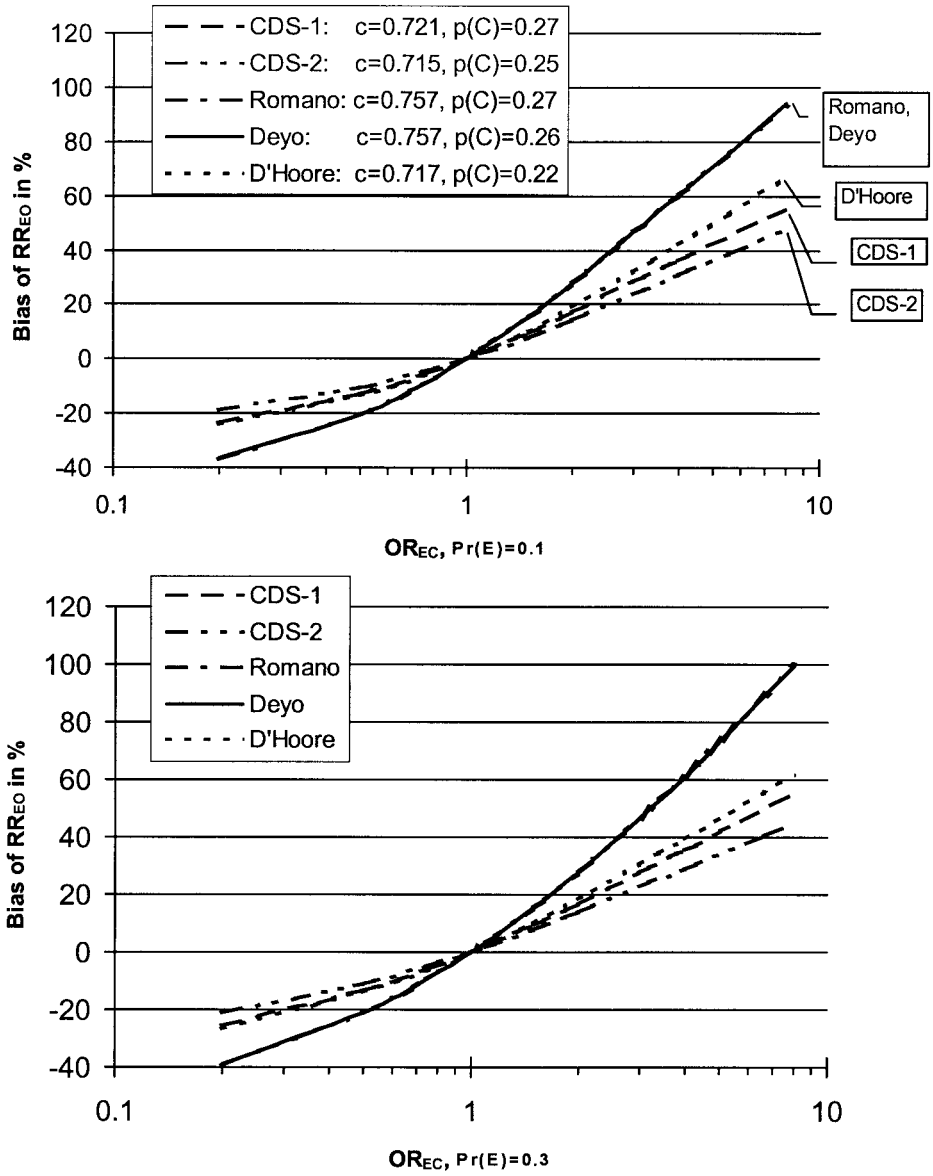

FIGURE 1. Percentage of relative bias, by comorbidity, that can be controlled by five comorbidity scores as a function of the exposure-comorbidity association OR_{EC} and the prevalence of exposure Pr(*E*). An exposure-mortality association of $RR_{EO} = 1$ was assumed. As shown, the Romano and Deyo scores completely overlap in both plots. The prevalence of comorbidities, Pr(*C*), varies according to the observed prevalences of the scores in a British Columbia, Canada, population aged ≥65 years that used antihypertensives, 1995–1997. RR, relative risk; E, exposure; O, outcome; CDS, Chronic Disease Score; c, *c* statistic; p, prevalence; C, confounder = comorbidity; OR, odds ratio. CDS-1, reference (7); CDS-2, reference (14); Romano, references (5, 6); Deyo, reference (8); D'Hoore, references (9, 10).

tinct drugs into number of chronic diseases involves loss of information on disease severity. Although the CDS considers multiple drug therapy versus monotherapy of heart disease and respiratory illness, it fails to do so for other diagnoses and does not account for medication changes as disease progresses.

Zhang et al. (35) suggested combining multiple Deyo scores based on ICD-9 diagnoses from different data sources, including hospital discharge, outpatient physician services, and auxiliary services (nursing facilities, home health aid, etc.), to improve performance. With a model that adjusted for age and gender, these authors reported a 3 percent improvement in the *c* statistic to predict mortality (0.702 to 0.724) in a random sample of Medicare enrollees. When we constructed a model that included the same set of covariates but without auxiliary information, we observed only a 1.5 percent improvement (0.757 to 0.768), which is closer to the 1.1 percent improvement observed in a recent study of breast cancer patients (36). Additional improvement (2 percent) was achieved when we combined the ICD-9-based score with the medication-based CDS-1 score. Since the combination that included number of distinct med-

ications received during the baseline year performed equally well, we suggest its combination with the Romano or Deyo score as an easily applicable and improved measure of comorbidity.

Our data support earlier findings (12) of almost no difference between modeling comorbidity scores as a continuous variable or as several categories. Binary coding is not recommended for D'Hoore's score, since it lost 2.7 percent of its *c* statistic when compared with a continuous model. Including quadratic terms of the scores makes interpretation of coefficients more difficult, with almost no gain in prediction.

In their original publication, Ghali et al. claimed that their score performed almost 15 percent better in predicting mortality than the Deyo score did ($c = 0.70$ vs. $c = 0.61$) (11). However, they empirically chose the weights of their abridged Charlson score to optimize prediction of mortality in their sample of patients with coronary bypass surgery. In our study of elderly recipients of antihypertensive medications, who constitute about one third of the total British Columbia population aged 65 years or more, generic scores such as those of Deyo or Romano performed better. This conclusion confirms earlier findings of Roos et al. (37) that performance of the Deyo score in predicting 1-year mortality can change considerably in specific disease groups, such as patients undergoing prostatectomy ($c = 0.64$), cholecystectomy ($c = 0.70$), or bypass surgery ($c = 0.75$).

Although the *c* statistics of the CDS-1 and Romano scores are statistically different, the question remains whether it is worthwhile to purchase and process diagnostic data in addition to pharmacy data to improve the *c* statistic from 0.738 to 0.783 (CDS-1 combined with Romano), an improvement of 9 percent in terms of the range between chance ($c = 0.5$) and perfect ($c = 1.0$) prediction. On the basis of detailed discharge data that included demographics and up to four comorbidities per patient, Hannan et al. (38) reported a *c* statistic of 0.742 for prediction of in-hospital mortality in patients with bypass surgery in New York State. After important clinical predictors were added, including ejection fraction, >90 percent narrowing of the left main vessel, and reoperation, the *c* statistic improved to 0.790, that is, 9.6 percent of the range from chance to perfect. Other authors (39) concluded that there is a significant difference between $c = 0.72$ and $c = 0.74$ in National Cholesterol Education Program guidelines I and II in predicting cardiovascular mortality. From this and other examples, it appears that large investments yield only small numeric gains in *c* statistics above 0.75. Whether those gains are worthwhile depends on the benefits of a "truer" analysis and the costs of error, which are unique to each problem.

In addition to measuring the relative predictive abilities of scores, we estimated their relative abilities to reduce confounding bias. Although our analyses of the effects on confounding bias relied on simplifying assumptions (e.g., dichotomous comorbidity measures and a single confounder), they suggest that more confounding could possibly be controlled by the Romano and Deyo scores than by the other scores.

The present study estimated and ranked the performance of six published comorbidity scores for a variety of endpoints in claims databases, but the generalizability of our results may be limited to an elderly, predominantly White population aged 65 years or more with equal access to state-funded health care. Performance of the Deyo score in the British Columbia population was better than in a random sample of Medicare enrollees (35). We caution against assuming performances will be similar in patient subgroups with specific diagnoses or of low-income (Medicaid) status. Relative performance depends on data quality. Similar studies of comparative performance are needed with other databases.

Although comorbidity scores are useful because they are easy to use and they save time and resources (a major issue when analyzing massive health care databases), they provide only a limited ability to control for confounding (1). Adjusting for a score should not be regarded as successfully controlling for confounding, because a summary score imposes on the analysis a fixed model of the relation between comorbidities and outcome, which is likely to differ among populations (40, 41). In addition, when the outcomes of a particular disease are studied, effects may be underestimated if the disease is a major ingredient of the score. If the goal is to control confounding as best as the data permit, scores are still useful for preliminary analyses to indicate the direction and magnitude of confounding, which can guide decisions about further analyses. The benefit versus the cost of using more thorough approaches to control confounding versus comorbidity scores is a topic that requires further research.

ACKNOWLEDGMENTS

Supported by the Drug Information Association, Fort Washington, Pennsylvania, and Pharmacare, Ministry of Health of British Columbia. Dr. Schneeweiss was supported by grants from the Deutsche Forschungsgemeinschaft (DFG#Schn527/3-1 and DFG#Schn527/4-1) and the US Agency for Healthcare Research and Quality (RO3 HSO9855 and RO1 HS10881) and by a Pharmacoepidemiology Training and Research Grant, Harvard University, Boston, Massachusetts.

REFERENCES

1. Schneeweiss S, Maclure M. Use of comorbidity scores for control of confounding in studies using administrative databases. Int J Epidemiol 2000;29:891–8.
2. Greenfield S, Nelson EC. Recent developments and future issues in the use of health status assessment measures in clinical settings. Med Care 1992;30(5 suppl):MS23–41.
3. Iezzoni LI. Risk adjustment for measuring healthcare outcomes. 2nd ed. Chicago, IL: Health Administration Press, 1997.
4. Roos LL, Sharp SM, Cohen MM, et al. Risk adjustment in claims-based research: the search for efficient approaches. J Clin Epidemiol 1989;42:1193–206.
5. Romano PS, Roos LL, Jollis JG. Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives. J Clin Epidemiol 1993;46:1075–9.

6. Romano PS, Roos LL, Jollis JG. Further evidence concerning the use of a clinical comorbidity index with ICD-9-CM administrative data. J Clin Epidemiol 1993;46:1085–90.
7. Von Korff M, Wagner EH, Saunders K. A chronic disease score from automated pharmacy data. J Clin Epidemiol 1992;45: 197–203.
8. Deyo RA, Cherkin DC, Ciol MA. Adapting a clinical comorbidity index for use with ICD-9-CM administrative databases. J Clin Epidemiol 1992;45:613–19.
9. D'Hoore W, Sicotte C, Tilquin C. Risk adjustment in outcome assessment: the Charlson comorbidity index. Methods Inf Med 1993;32:382–7.
10. D'Hoore W, Bouckaert A, Tilquin C. Practical considerations on the use of the Charlson index with administrative data bases. J Clin Epidemiol 1996;49:1429–33.
11. Ghali WA, Hall RE, Rosen AK, et al. Searching for an improved clinical comorbidity index for use with ICD-9-CM administrative data. J Clin Epidemiol 1996;49:273–8.
12. Melfi C, Holleman E, Arthur D, et al. Selecting a patient characteristics index for the prediction of medical outcomes using administrative claims data. J Clin Epidemiol 1995;48:917–26.
13. Poses RM, Smith WR, McClish DK, et al. Controlling for confounding by indication for treatment. Are administrative data equivalent to clinical data? Med Care 1995;33:AS36–AS46.
14. Clark DO, von Korff M, Saunders K, et al. A chronic disease score with empirically derived weights. Med Care 1995;33: 783–95.
15. Schneeweiss S, Soumerai SB, Glynn RJ, et al. Intended and unintended impacts on antihypertensive drug use by a policy of differential cost sharing for angiotensin-converting enzyme inhibitors. Pharmacoepidemiol Drug Safety 2000;9:S65.
16. Charlson ME, Pompei P, Ales KL, et al. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis 1987;40:373–83.
17. Charlson ME, Szatrowski TP, Peterson J, et al. Validation of a combined comorbidity index. J Clin Epidemiol 1994;47: 1245–51.
18. AHFS drug information 96. Bethesda, MD: American Society of Health-System Pharmacists, 1996.
19. Diehr P, Yanez D, Ash A, et al. Methods for analyzing health care utilization and costs. Annu Rev Public Health 1999;20: 125–44.
20. British Columbia migration—outlook for 2000. Victoria, British Columbia, Canada: BC Stats, 1999. (http://www.bcstats.gov.bc.ca/pubs/pr_mig.htm#mig).
21. Anderson GM, Kerluke KJ, Pulcins IR, et al. Trends and determinants of prescription drug expenditures in the elderly: data from the British Columbia Pharmacare Program. Inquiry 1993;30:199–207.
22. Williams JI, Young W. Inventory of studies on the accuracy of Canadian health administrative databases. Halifax, Nova Scotia, Canada: Population Health Research Unit, Dalhousie University, Institute for Clinical Evaluative Sciences, 1996. (Publication no. 96-03-TR).
23. Fowles JB, Lawthers AG, Weiner JP, et al. Agreement between physicians' office records and Medicare Part B claims data. Health Care Financ Rev 1995;16:189–99.
24. Romano PS, Mark DH. Bias in the coding of hospital discharge data and its implications for quality assessment. Med Care 1994;32:81–90.
25. Glynn RJ, Monane M, Gurwitz JH, et al. Agreement between drug treatment data and a discharge diagnosis of diabetes mellitus in the elderly. Am J Epidemiol 1999;149:541–9.
26. Fisher ES, Whaley FS, Krushat WM, et al. The accuracy of Medicare's hospital claims data: progress has been made but problems remain. Am J Public Health 1992;82:243–8.
27. Ash AS, Shwartz M. Evaluating the performance of risk-adjustment methods: dichotomous outcomes. In: Iezzoni LI, ed. Risk adjustment for measuring healthcare outcomes. 2nd ed. Chicago, IL: Health Administration Press, 1997.
28. Wilson PWF, D'Agostino RB, Levy D, et al. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97:1837–47.
29. Hosmer DW, Lemeshow S. Applied logistic regression. 2nd ed. New York, NY: John Wiley & Sons, 2000.
30. Liebetrau AM. Measures of association, quantitative application in the social sciences. Vol 32. Beverly Hills, CA: Sage Publications, 1983.
31. Shwartz M, Ash AS. Evaluating the performance of risk-adjustment methods: continuous outcomes. In: Iezzoni LI, ed. Risk adjustment for measuring healthcare outcomes. 2nd ed. Chicago, IL: Health Administration Press, 1997.
32. Dudley RA, Harrell FE, Smith LR, et al. Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary bypass graft surgery. J Clin Epidemiol 1993; 46:261–71.
33. Glynn RJ, Monane M, Gurwitz JH, et al. Aging, comorbidity, and reduced rates of drug treatment for diabetes mellitus. J Clin Epidemiol 1999;52:781–90.
34. Redelmeier DA, Tan SH, Booth GL. The treatment of unrelated disorders in patients with chronic medical diseases. N Engl J Med 1998;338:1516–20.
35. Zhang JX, Iwashyna TJ, Christakis NA. The performance of different lookback periods and sources of information for Charlson comorbidity adjustment in Medicare claims. Med Care 1999;37:1128–39.
36. Wang PS, Walker A, Tsuang M, et al. Strategies for improving comorbidity measures based on Medicare and Medicaid claims data. J Clin Epidemiol 2000;53:571–8.
37. Roos LL, Stranc L, James RC, et al. Complications, comorbidities, and mortality: improving classification and prediction. Health Serv Res 1997;32:229–38.
38. Hannan EL, Kilburn H, Lindsey ML, et al. Clinical versus administrative data bases for CABG surgery. Does it matter? Med Care 1992;30:892–907.
39. Grover S, Coupal L, Hu XP. Identifying adults at increased risk of coronary disease: how well do the current cholesterol guidelines work? JAMA 1995;274:801–6.
40. Michels KB, Greenland S, Rosner BA. Does body mass index adequately capture the relation of body composition and body size to health outcomes? Am J Epidemiol 1998;147:167–72.
41. Katz D, Foxman B. How well do prediction equations predict? Using receiver operating characteristic curves and accuracy curves to compare validity and generalizability. Epidemiology 1993;4:319–26.
42. Walker AM. Observation and inference. An introduction to the methods of epidemiology. Newton Lower Falls, MA: ERI, 1991.

APPENDIX

We derived an equation that relates the apparent relative risk of an exposure-outcome association in the presence of confounding to the associations between confounder and exposure as well as confounder and outcome.

Assuming a 2-by-2 table of a dichotomous exposure and a dichotomous confounder, let e be the prevalence of exposed patients with the confounder present. The association between confounder and exposure can then be measured by the confounder-exposure odds ratio or OR_{CE}, which is a function of e and the marginal probabilities of exposure $\Pr(E)$ and confounder $\Pr(C)$ (e.g., Walker (42)):

$$OR_{CE} = \frac{e[1 - \Pr(C) - \Pr(E) + e]}{[\Pr(C) - e][\Pr(E) - e]}. \quad (1)$$

Assuming no underlying true exposure-outcome association

or $RR_{EO} = 1$, Walker (42) showed that the apparent or crude RR_{EO} is a function of e, the marginal probabilities $Pr(E)$ and $Pr(C)$, and the confounder-outcome association RR_{CO}:

$$\text{crude } RR_{EO} = \frac{e[RR_{CO} - 1] + Pr(E)}{[Pr(C) - e][RR_{CO} - 1] - Pr(E) + 1} \frac{1 - Pr(E)}{Pr(E)}. \quad (2)$$

Solving equation 1 for e

$$\underbrace{e^2(OR_{CE} - 1)}_{a} + \underbrace{e[- Pr(C)OR_{CE} - Pr(E)OR_{CE} + Pr(E) + Pr(C) - 1]}_{b} + \underbrace{Pr(C)OR_{CE}Pr(E)}_{c} = 0,$$

e can be found as the solution to a quadratic equation.

Substituting the derived term for e in equation 2 yields the crude RR_{EO} as a function of OR_{CE}, RR_{EO}, RR_{CO}, and the marginal probabilities $Pr(E)$ and $Pr(C)$.