

Use of comorbidity scores for control of confounding in studies using administrative databases

Sebastian Schneeweiss^{a,b} and Malcolm Maclure^{a,c}

Background	Comorbidity scores are increasingly used to reduce potential confounding in epidemiological research. Our objective was to compare metrical and practical properties of published comorbidity scores for use in epidemiological research with administrative databases.
Methods	The literature was searched for studies of the validity of comorbidity scores as predictors of mortality and health service use, as measured by change in the area under the receiver operating characteristic (ROC) curve for dichotomous outcomes, and change in R^2 for continuous outcomes.
Results	Six scores were identified, including four versions of the Charlson Index (CI) which use either the three-digit International Classification of Diseases, Ninth Revision (ICD-9) or the full ICD-9-CM (clinical modification) code, and two versions of the Chronic Disease Score (CDS) which used outpatient pharmacy records. Depending on the population and exposure under study, predictive validities varied between $c = 0.64$ and $c = 0.77$ for in-hospital or 30-day mortality. This is only a slight improvement over age adjustment. In one study the simple measure 'number of diagnoses' outperformed the CI ($c = 0.73$ versus $c = 0.65$). Proprietary scores like Ambulatory Diagnosis Groups and Patient Management Categories do not necessarily perform better in predicting mortality. Comorbidity indices are susceptible to a variety of coding errors.
Conclusions	Comorbidity scores, particularly the CDS or D'Hoore's CI based on three-digit ICD-9 codes, may be useful in exploratory data analysis. However, residual confounding by comorbidity is inevitable, given how these scores are derived. How much residual confounding usually remains is something that future studies of comorbidity scores should examine. In any given study, better control for confounding can be achieved by deriving study-specific weights, to aggregate comorbidities into groups with similar relative risks of the outcomes of interest.
Keywords	Comorbidity, confounding, risk adjustment, health services epidemiology, clinical epidemiology
Accepted	13 March 2000

Health status, as measured by disease history, has long been recognized as a major class of potential confounder in most epidemiological studies. Over the last two decades, a variety of methods have been developed that might permit more uniform

comorbidity adjustment across different epidemiological studies.^{1,2} Single scores for summarizing comorbidity are of particular interest to investigators who use very large administrative databases because the first stage of such studies is to digest a vast array of administrative variables, some of which have long lists of possible values, into an intelligible and manageable set of proxy variables. In this process, the benefits of simplification usually seem much greater than the risks of oversimplification.

The simplest comorbidity score is also the most widely used measure of confounding in epidemiology: age. Although it is a relatively poor index of comorbidity, it is recorded accurately and ubiquitously in administrative databases, and methods of

^a Department of Epidemiology, Harvard School of Public Health, Boston, MA, USA.

^b Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA.

^c Pharmacare, Ministry of Health, British Columbia, Canada.

Reprint requests to: Sebastian Schneeweiss, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, 221 Longwood Ave, Boston, MA 02115, USA. E-mail: sschneew@hsph.harvard.edu

adjusting for age are standard. It is known to be a usually necessary and often insufficient adjustment in most non-experimental studies. We began this review with the question whether any comorbidity score would improve on adjustment for age, while remaining almost as simple.

On first thought, there would seem to be at least two major advantages of a perfect single comorbidity score, if such a score could be created. (1) in multivariate modelling, a single score summarizing comorbidity would increase the statistical efficiency of analysis, compared with modelling each individual morbidity separately, especially when other risk factors modify the effects of comorbidity, necessitating use of two- or three-way product terms; (2) a validated comorbidity instrument available as a standard off-the-shelf product would simplify the process of variable selection both in the design and analysis of a study, and might increase the comparability of findings from different studies.

On second thought, after an initial look at literature on the difficulty of adjusting for comorbidity in administrative datasets,^{3,4} we developed the impression that these advantages might be distant or unattainable. A prominent example of controversy over comorbidity adjustment was the claim by Roos *et al.* that, in contrast to suprapubic prostatectomy for benign prostatic hyperplasia, transurethral prostatectomy appeared to be associated with increased 5-year mortality after comorbidity adjustment based on discharge information (RR = 1.45, 95% CI: 1.15–1.83).^{5,6} Concato *et al.* repeated the study and found similarly elevated risks with the same adjustment method.⁷ However, the increase in mortality vanished (RR = 1.03, 95% CI: 0.51–2.07) when the same method was used, however, based on medical record review. The authors concluded that reports of comorbidity-adjusted results from automated databases might be insufficient, tending to underrepresent some comorbid conditions⁸ and should be interpreted cautiously.

Chart review is rarely possible in the growing number of low-budget studies using administrative data. Therefore, comorbidity adjustment based on administrative data with all its limitations is here to stay. To proceed with such research and avoid erroneous inferences requires a mixture of scepticism about the ability of comorbidity scores to fully control for confounding, and cautious optimism that comorbidity scores are nevertheless useful tools in hypothesis-screening studies, provided they have been validated. To achieve the right mixture of scepticism and optimism, we need to understand the evidence from validation studies of existing comorbidity scores.

Assessing metrical properties

Validity

There is no gold standard for 'true comorbidity'. The construct 'comorbidity' is a complex function of the number and severity of comorbidities in relation to the primary diagnosis. Because of the absence of a gold standard, researchers use the assumption that 'true comorbidity' is correlated with worse health outcomes, health care utilization, and costs. Therefore, the validity of a comorbidity index is assessed by how well the index predicts those outcomes which indirectly determines how well comorbidity indices can control for confounding.

For continuous outcomes the most common measure of validity is the improvement in explained variance, R^2 , of a

linear regression model after adding the score to some baseline model— R^2 ranges from 0 to 1 with increasing explained variance. For dichotomous outcomes, there are measures of discrimination and measures of calibration. Measures of discrimination compare the predicted outcome with the actual outcome, e.g. the *c* statistic, which is equivalent to the area under curve (AUC) of a receiver operating characteristic (ROC). The AUC or *c* range from 0 to 1 with 1 indicating a perfect prediction and 0.5 indicating a chance prediction. Measures of calibration, like the Hosmer-Lemeshow goodness-of-fit statistic (H-L), assess the discrimination in all strata of the index values separately and summarize the information.⁸ The statistic or the corresponding *P*-value are usually reported. In early papers in particular, the validity of prediction is often assessed by the strength of an association between the comorbidity index and the outcome, in terms of the odds ratios (OR) or relative risk (RR) per increment in score.

Reliability

Computation of an index itself is completely reproducible in the same set of recorded data. Computerized indices from administrative databases are in that sense completely reliable but depend on the accuracy of information stored in the database. Administrative data are derived from what is documented in the medical or pharmacy record. Therefore, the real focus of reliability for code-based measures involves how accurately and completely the coded information was gathered. Thus reliability of code-based comorbidity indices is usually not tested directly but only inferred from reports of other investigators addressing coding accuracy.¹ We will raise issues of data accuracy in a separate section.

Structural characteristics and predictive validity

Populations, settings and prognostic endpoints

We identified original research on the metric properties of comorbidity indices by an extensive literature search using Medline and HealthStar databases, as well as bibliographies and expert consultations. We identified six distinct indices of comorbidity for use in administrative databases which were published with at least some analyses of their predictive performance (Table 1).^{10–18} We excluded systems that do not combine conditions into a single summary score but consist of multiple categorical (e.g. Elixhauser *et al.*¹⁹) or binary variables (Ambulatory Care Groups²⁰) for each diagnostic group. Those systems might control confounding slightly better, however, a regression model that uses 30 or 32 variables to control for comorbidity limits the ability to model interactions and decreases precision in epidemiological analyses.

The six scores were tested in different settings in North America. Four of the six use diagnostic information from ICD-9 codes and are based on the Charlson Index (Table 2)²¹ and two scores are based on the use of prescription medications as indicators of chronic comorbid conditions (Table 3). The statistical properties of some instruments were explored in specific populations, including patients after bypass surgery, lumbar disc surgery, or with ischaemic heart disease. Prognostic endpoints vary, including one-year mortality, in-hospital mortality, length of hospital stay, re-admissions, future hospitalizations, ambulatory visits, or resource use (Tables 2 and 3).

Table 1 Comorbidity measuring instruments for use in administrative databases

Year	Name of instrument or author(s) of first publication	Short title	Data basis
1989/1993/1997	Dartmouth-Manitoba method ^{10-12,18}	DM-CI	ICD-9-CM
1992	Chronic Disease Score ¹³	CDS	AHFS ^a (4 digits)
1993	Deyo <i>et al.</i> ¹⁴	Deyo-CI	ICD-9-CM
1993	D'Hoore <i>et al.</i> ^{15,16}	D'Hoore-CI	ICD-9 (3 digits)
1995	Extended Chronic Disease Score ²⁵	CDS-2	AHFS
1996	Ghali <i>et al.</i> ¹⁷	Ghali-CI	ICD-9-CM

^a American Hospital Formulary Services based on prescription drug dispensing.

Table 2 Charlson Index based comorbidity measures

	DM-CI Dartmouth-Manitoba code 1989/1993	Deyo-CI Deyo <i>et al.</i> 1993	D'Hoore-CI D'Hoore <i>et al.</i> 1993/1996	Ghali-CI Ghali <i>et al.</i> 1996
Setting of original study	(a) Hospital discharges from a Manitoba (Canada) hospital April 1980–May 1992. (b) Discharges from licensed federal hospitals in California (USA) 1988–1991.	Medicare beneficiaries in the USA in 1989.	792 839 records in 78 hospitals with > 100 beds in Quebec (Canada) in the administrative year 1989/90.	524 740 hospital discharges of Massachusetts (USA) residents from all MA hospitals in 1990 (development sample) and 1992 (test sample).
Information used	(a) From Manitoba Health Services Commission files: Up to 16 ICD-9-CM coded diagnoses of index admission and a 12-month period before index. (b) From Hospital Discharge Data Program discharge abstracts: ICD-9-CM coded primary diagnoses and up to 24 secondary.	From Medicare database: ICD-9-CM coded diagnoses during hospital stay with and without diagnoses the year before index.	From MED-ECHO database: First three digits of ICD-9 coded principal diagnosis and 15 secondary diagnoses.	From Massachusetts Health Data Consortium discharge abstracts: Up to 15 ICD-9-CM coded diagnoses.
Population under study	(a) 4121 patients who underwent bypass surgery aged ≥30 years. (b) 55 407 patients who underwent lumbar disc excision surgery aged ≥18.	27 111 patients who underwent lumbar spine surgery with a mean age of 72 years.	33 940 hospitalized patients with ischaemic heart disease with a mean age of 63 years.	6326 patients who underwent bypass surgery with a mean age of 65 years.
Prognostic endpoint	One-year mortality, significant in-hospital complications, and 90-day re-admission.	Post-op death (in-hospital or 6 weeks after discharge), post-op complications, length of stay, hospital charges.	In-hospital mortality	In-hospital mortality
Summarization	Charlson Index	Charlson Index	Charlson Index	Charlson Index

Table 3 Comorbidity measures not based on the Charlson Index

	CDS Von Korff <i>et al.</i> 1992	CDS-2 Clark <i>et al.</i> 1995
Setting of original study	Enrollees of the Group Health Cooperative of Puget Sound in Seattle and surrounding communities in 1985.	Enrollees of the Group Health Cooperative of Puget Sound in Seattle and surrounding communities in 1992.
Information used	From Group Health Cooperative pharmacy database: information on all prescription medications filled in a HMO pharmacy (90% of all fills of enrollees).	From Group Health Cooperative pharmacy database: all prescription medications coded with AHFS ^a and filled in a HMO pharmacy (90% of all fills of enrollees).
Population under study	All 122 911 adult enrollees in Central and East region.	Development sample: randomly 50% of adult enrollees: 125 000. Test sample: second half of adult enrollees.
Prognostic endpoint	One-year mortality, one-year hospitalization rate	One-year primary care visits, one-year total costs, one-year outpatient costs
Summarization	Chronic Disease Score (CDS) summarizing the patterns of selected prescription medications usage during a one-year period. Weights are derived by expert panels.	CDS-2 summarizing an extended list of medications used over a 6-month-period. Weights are derived empirically.

^a American Hospital Formulary Services.

Charlson Index based indices

The Charlson Index (CI) is a list of 19 conditions (Table 4).²¹ Each condition has a weight assigned from 1 to 6, which was derived from relative risk estimates of a proportional hazard regression model using clinical data. The CI score is the sum of weights for all prevalent conditions. Although the index might seem rather crude, it showed a strong monotonic association of a 2.3-fold (95% CI: 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,²¹ and similar results for post-operative survival in patients with hypertension or diabetes.²²

Deyo *et al.*

Deyo *et al.*¹⁴ assigned ICD-9-CM codes to all 19 diagnoses in the Charlson Index and assessed the index in a study of 27 111 Medicare beneficiaries (age >65) who underwent lumbar spine surgery. They reported statistically significant associations of Charlson scores with increased number of blood transfusions, length of stay, hospital charges, discharges to nursing homes, and 6-week post-operative mortality in an analysis of variances ($P < 0.0005$). The Deyo-CI was associated with hospital complications with $P < 0.01$. Magnitudes of associations were not reported. Similar results could be observed when only information on current hospital stay was used in contrast to diagnostic information on any hospitalization in the preceding year. The authors claim that these associations were significant 'even after controlling for patient age' which means that controlling for CI in addition to age could improve the results of a risk adjustment procedure as compared to adjusting for age only. The original paper does not show how much variation is explained by any outcome.

Melfi *et al.*²³ used the Deyo-CI in studying 249 744 Medicare patients who underwent total knee replacement between 1985

and 1989. An increase in the Deyo-CI of one point increased the probability of 30-day post-operative mortality by 17%. The Deyo-CI model showed a c statistic of 0.653 in predicting 30-day post-operative mortality (baseline model = 0.645) and an R^2 of 0.175 in predicting length of stay (baseline model = 0.174). The improvement in predictive power of models including Deyo-CI is marginal, mostly because the 'baseline' model already consists of many known important predictors: gender, age, race, socio-economic status, area of residence and level of care to which the patient was discharged. The same model with Deyo-CI substituted by the number of distinct diagnoses shows a better prediction of mortality ($c = 0.73$).

Poses *et al.*²⁴ used Deyo's index to predict in-hospital death at a university hospital. They reported areas under the ROC curve of $c = 0.64$ for the Deyo-CI, $c = 0.62$ for duration in hospital, and $c = 0.61$ for age. The ICD-based Charlson Index contributed in a multivariate model to the prediction of in-hospital death, even when clinical data were added (OR = 1.2, 95% CI: 1.1–1.4).

Dartmouth-Manitoba

The Dartmouth-Manitoba version of the Charlson Index (DM-CI) by Roos *et al.*^{10,11} was the first adaptation of the Charlson Index to administrative databases. It includes additional conceptually similar ICD diagnoses which were not explicitly listed in Charlson's original list of 19. This was meant to increase the sensitivity of the DM-CI. In a study of one-year mortality following several types of surgery, a multivariate analysis controlling for age and sex showed that the DM-CI did contribute significantly to the fit for prostatectomy and bypass surgery but not in cholecystectomy.

Ghali *et al.*¹⁷ assessed the agreement between scores from DM-CI and Deyo-CI in a population of 6791 bypass surgery patients. They found 90% of patients were assigned identical scores by the two methods and the two scores differed by only 1 among a further 5% of patients.

Romano *et al.*¹² compared RR estimates in the presence of 16 individual comorbid conditions that are part of the DM-CI and Deyo-CI in the same bypass patients ($n = 4121$) and in patients with lumbar discectomy ($n = 55\ 296$). They show that individual diagnoses defined by Deyo-CI and DM-CI methods have similarly RR estimates within the same population (bypass respectively discectomy patients) and the same outcome (death respectively complication). However, the RR changed strongly between populations, e.g. RR = 1.5 for metastatic tumour in bypass population to predict mortality, RR = 4.4 for metastatic tumour in lumbar discectomy patients to predict complication. They provided no direct quantitative comparisons of the two scores.

In 1997, Roos *et al.* published¹⁸ an augmented version of DM-CI which added coagulopathy, neurological disorders, hypertension, arrhythmia, pneumonia and malnutrition. The predictive validity of models, already controlling for age and gender, improved from $c = 0.64$ to $c = 0.68$ for bypass surgery, from $c = 0.70$ to $c = 0.77$ for pacemaker surgery, and from $c = 0.75$ to $c = 0.76$ for hip fracture repair.

Ghali *et al.*

Rather than adding diagnoses, Ghali *et al.*¹⁷ reduced the number of diagnoses by selecting those that best predicted in-hospital mortality among 6326 bypass surgery patients identified from 257 333 Massachusetts hospital discharges. Weights

Table 4 Weighted index of comorbidity according to Charlson *et al.*²¹ and Ghali *et al.*¹⁷

Charlson weights	Conditions	Ghali weights
1 ^a	Myocardial infarct	1 ^a
1	Congestive heart failure	4
1	Peripheral vascular disease	2
1	Cerebrovascular disease	1
1	Dementia	–
1	Chronic pulmonary disease	–
1	Connective tissue disease	–
1	Ulcer disease	–
1	Mild liver disease	–
1	Diabetes	–
2	Hemiplegia	–
2	Moderate or severe renal disease	3
2	Diabetes with end organ damage	–
2	Any tumour	–
2	Leukaemia	–
2	Lymphoma	–
3	Moderate or severe liver disease	–
6	Metastatic solid tumour	–
6	AIDS	–

^a Charlson: acute and old myocardial infarct (MI); Ghali: acute MI = 1, old MI = 0.

were changed in order to further improve the predictive performance (Table 4). The Ghali-CI predicted in-hospital mortality in a similar population of bypass patients one year later slightly better than the DM-CI. The Ghali-CI had discrimination measures of $c = 0.74$ and $R^2 = 0.034$ compared with $c = 0.704$ and $R^2 = 0.018$ for Deyo-CI. This relatively small gain is noteworthy because the weights for the Ghali-CI were derived from a similar group of bypass patients, whereas those for the Deyo-CI were based on the original CI developed from a study of breast cancer patients using 19 comorbidities.

D'Hoore *et al.*

Motivated by the fact that some institutions, frequently outside the US, use only ICD-9 codes without the Clinical Modification (CM), and coding of the trailing digits in the ICD-9-CM is less reliable, D'Hoore *et al.* designed a Charlson Index (D'Hoore-CI) using only the first three digits of ICD-9.¹⁵ Using the MED-ECHO database restricted to 78 hospitals with 792 839 discharges (Table 2), they evaluated the prediction of in-hospital mortality among 33 940 patients with a principal diagnosis of ischaemic heart disease.¹⁶ Using a model with age, sex, and acute myocardial infarction as the principal diagnosis, the D'Hoore-CI showed good predictive performance for two consecutive 2-year periods in the same population: 1989/90: $c = 0.87$, and $R^2 = 0.14$; 1990/91: $c = 0.86$, and $R^2 = 0.13$.¹⁶ Discrimination for other primary diagnoses was reported: ischaemic heart disease ($c = 0.81$), congestive heart failure ($c = 0.67$), stroke ($c = 0.66$), and bacterial pneumonia ($c = 0.82$).¹⁵ Primary diagnoses were not excluded from the Charlson Index.

Comorbidity scores based on outpatient pharmacy data

The Chronic Disease Score (CDS) uses pharmacy dispensing data to assign patients to chronic disease groups. An integer weight is given to each comorbidity group represented by selected medication classes, which are summed to the overall score.¹³ The CDS was developed using the judgement of an interdisciplinary expert group of researchers and practitioners. In several pilot studies with varying populations within the Group Health Cooperation Health Plan of Puget Sound (GHC), the derived CDS was compared to clinical judgement and standard instruments to measure self-rated health status, psychological impairment, chronic pain status, and functional disability. The CDS was eventually tested among all 122 911 GHC enrollees. A multivariate logistic regression model showed that, with an increasing CDS score, the probabilities of one-year hospitalization and one-year mortality steadily increased. The highest CDS score category (7+) had a 10 times higher probability of dying the next year than CDS = 0. This effect diminished by up to 50% when age, gender and number of ambulatory visits were included in the prediction model.

An extended version, CDS-2,²⁵ used a 50% development sample out of 250 000 managed care enrollees to derive empirical weights from a multiple logistic regression model. The CDS-2 was tested in the remaining sample of 125 000 enrollees of the same insurance plan of the same year together with the original CDS. The authors claimed that CDS-2 had a stronger association with one-year mortality, one-year hospitalization and health care utilization although no original data were shown. It remains unclear to what extent the improved prediction is due to the close similarity of the development and test sample or to an

intrinsically better performance. In their study, the proportion of the explained variance in predicting costs and health care utilization was reported for the CDS and CDS-2, making both instruments directly comparable (Table 5). The CDS showed correlations with the SF-36 health status instrument and the BSI-8 depression disorder screening instrument.²⁶

Table 5 summarizes the predictive performance of all instruments. Direct comparisons of indices are recommended only within the same studies (grey background). It is important to note that proprietary products such as the ADG perform no better than CDS and the Patient Severity Score (PSL) and Relative Severity Score (RIS) scores of the Patient Management Categories perform only one percentage point better than a measure as simple as the number of distinct diagnoses during the last 6 months.²³ All studies modelled comorbidity indices as one ordinal variable. Modelling Deyo-CI as one ordinal variable compared to 3 or 13 categories changed R^2 values only in the third decimal.²³

Data accuracy

Coding

Diagnoses are occasionally erroneous, often coded incorrectly, and frequently omitted from administrative data.^{27,28} Romano *et al.* compared original DM-CI with ICD diagnoses coded by a trained abstractor from medical records to ICD codes from claims data in 1067 patients.¹² Administrative data tended to underestimate comorbid conditions. Record-based scores of 0, 1, 2, and 3+ were observed in 38%, 32%, 18% and 12% of the patients, in contrast to 65%, 27%, 7%, and 1% in the DM-CI based on administrative data. In a comparison of four populations in different regions of the US²⁹ with either claims data or ICD codes derived from medical records, striking differences were found in the prevalence of comorbid conditions between ICD codes derived from records and those derived from claims data. In clinical data, histories of myocardial infarct and hyperlipidaemia were found more often. However, other major comorbidities, like diabetes and congestive heart failure, were nearly equally prevalent. It remains unclear to what degree the variation is due to differences in populations or methods. Similar results were reported in a comparison of 485 claims-based DM-CI indices versus chart-based DM-CI (overall Kappa = 0.36).⁸ Diabetes, tumours, and cirrhosis showed high agreement between reporting method (Kappa equal to 0.86, 0.78, and 0.67, respectively) but dementia, paralysis, and ulcer showed weak agreement (Kappa equal to 0.29, 0.19, and 0.14, respectively). Less optimistic agreements were reported for patients undergoing carotid endarterectomy comparing ICD diagnoses coded from records with Medicare data. Kappa statistics for diabetes and tumours equalled 0.68 and 0.20.³⁰

Complications and diagnostic examinations

Another problem, particularly in studies of in-hospital mortality, is that it is often not clear whether some diagnoses are comorbidities at hospital admission or complications during the hospital stay. Treating complications as comorbidities can result in an overoptimistic interpretation of an index's predictive capability for unfavourable outcomes. Roos *et al.*, however, showed that the impact of misinterpreting complications as comorbidities on the Charlson index is minor in surgical procedures.¹⁸

Table 5 Comparison of measures to assess the predictive validity of comorbidity measures used in administrative databases. Generally, measures should only be compared within each shaded box

Comorbidity indices and other predictors	Publication, population, and outcomes									
	Clark <i>et al.</i> ²⁵ Group Health Cooperative of Puget Sound			McFli <i>et al.</i> ²³ Medicare patients with total knee replacement		Poses <i>et al.</i> ²⁴ Medical service of onc University Hospital, USA	Ghali <i>et al.</i> ¹⁷ Bypass surgery hospital discharges in Massa- chusetts	D'Hoore <i>et al.</i> ¹⁵ Hospital discharges in Quebec	Roos <i>et al.</i> ¹⁸ Surgical hospital discharges from Manitoba hospitals	
	Total costs	Outpat costs	No of visits	LOS	30-day mortality	In-hospital mortality	In-hospital mortality	In-hospital mortality	In-hospital mortality	One-year mortality
Age + sex +	0.93			0.00		0.09				
+ CDS	0.97			0.18		0.10				
+ CDS-2	0.10			0.20		0.10				
+ ADG ^a	0.08			0.21		0.10				
+ ADG + CDS	0.10			0.24		0.18				
+ ADG + CDS-2	0.12			0.25		0.13				
Age + sex + race + region +				0.17		0.64				
+ PSL ^b				0.22		0.65				
+ RIS ^c				0.22		0.80				
+ No. of diagnoses (≤5)				0.21		0.75				
+ Deyo-CI				0.12		0.65				
Age						0.61				
Length of stay						0.62				
Deyo-CI						0.64				
APACHE II^d						0.83				
Age + sex +								0.66		
+ Deyo-CI								0.70		
+ Ghali _{score}								0.74		
+ Ghali _{empirical}								0.76		
Age + sex +										
+ D'Hoore-CI								0.76		
+ IHD + D'Hoore-CI ^e								0.81		
+ AMI + D'Hoore-CI								0.87		
+ Stroke + D'Hoore-CI								0.86		
+ BP + D'Hoore-CI								0.82		
+ CHF + D'Hoore-CI								0.76		
Age + sex +										
+ Prostatect + DM-CI										0.94
+ Prostatect + DM _A -CI										0.94
+ Cholecyste. + DM-CI										0.94
+ Cholecyste. + DM _A -CI										0.94
+ Bypass + DM-CI										0.94
+ Bypass + DM _A -CI										0.94

^a ADG = Ambulatory Diagnostic Group.

^b PSL = Patient Severity Score, a complex index derived from Patient Management Categories (PMC).

^c RIS = Relative Severity Score, a complex index derived from Patient Management Categories (PMC).

^d APACHE = Acute Physiology, Age, Chronic Health Evaluation, a complex score requiring detailed clinical information.

^e IHD = ischaemic heart disease, AMI = acute myocardial infarction, BP = bacterial pneumonia, CHF = congestive heart failure.

Diagnoses are sometimes recorded as present when the actual health service was to rule out that diagnosis. This is often distinguishable only indirectly with longitudinal data on the following encounters and procedures.

Completeness

Inaccuracies can occur when diagnoses are omitted because the data fields have been exhausted by more important diagnoses. Romano *et al.* showed that sensitivity to capture specific diagnoses

in administrative databases with five diagnosis fields reduced by an average 13 per cent points compared to a record with 25 fields.²⁸ The specificity is almost equal.

Prescription drugs as proxies for diagnoses can face reduced validity because they often have mixed indications, and because of a tendency to avoid prescribing additional drugs to patients who are already taking several and to reduce preventive medication in sicker patients.^{31,32}

Discussion

These studies of comorbidity scores suggest that they provide only a modest improvement on age adjustment. Why should such a mediocre comorbidity surrogate as age do well compared with comorbidity indices that have much higher face validity?

One major reason is that scores of any kind perform badly because they summarize a complex construct in an over-simplistic way making erroneous assumptions. Scores may perform reasonably well in the setting they were designed for and worse in other settings because they fail to represent a more general construct of comorbidity. The fact that only counting the number of distinct diagnoses performs equally well in one study demonstrates this but cannot be generalized.

Even if it provides only a modest improvement in ability to control for confounding, a simple score still would be appealing if, like age, it had excellent data accuracy and completeness and was widely used and understood. Unfortunately the accuracy of diagnostic data varies among databases, depending on the financial incentives and disincentives for recording it and the processes for error checking. The quality of data on drug use is less variable between databases, but drug use varies not only by disease status but also by ability to pay, prescribing customs and patient attitudes, which vary among health systems and regions.

Nevertheless we conclude that an off-the-shelf comorbidity score can still be a useful tool for exploratory data analysis. It enables one to assess the existence and direction of confounding by comorbidity. Although the magnitude of confounding is inevitably underestimated with such a score, it provides a preliminary quantification, which can guide the development of weights tailored to the particular study. For practical reasons, D'Hoore's¹⁵ adaptation of the Charlson Index, which uses only the first three digits of ICD-9 could be a starting point. If reliable pharmacy data are available, von Korff's Chronic Disease Scores^{13,25} might be an alternative with reasonable predictive validity for selected outcomes.

With this more modest purpose in mind, several criticisms of comorbidity scores are avoided. A major criticism of all comorbidity adjustment methods is that they were developed to predict a particular type of outcome (e.g. morbidity or mortality) but are used to adjust for the risk of other outcomes (e.g. health service use or costs). Although there are common cases when the severity of one outcome is inversely related to the intensity of health care use (e.g. sudden cardiac death), in the aggregate, adverse outcomes are positively correlated with each other. Therefore, a comorbidity score developed in one setting can be applied in a very different setting as long as it is only for exploratory purposes.

A major criticism of summary scores in general, including such diverse summaries as body mass index,³³ study quality

scores in meta-analyses,³⁴ and even age as a continuous variable, is that the summarization into a single value forces a relationship on the data that may be unrealistic. Even if the original Charlson weights, derived from regression coefficients predicting survival in patients at Cornell Medical Center had been the optimal fit for those data, it is unlikely it would fit the data as well in another population or for other outcomes. An alternative is to test alternative weights and to include several indicator terms in the model for different dimensions or categories of the variable. The numerically most efficient way of doing so is to model the outcome as a function of all comorbidity information including interaction terms and use regression coefficients to weight individual items of the scores in the same study. This criticism applies much more to final rigorous analyses than to initial exploratory data analysis. Although only supported by one published study, it may be not necessary to categorize scores, as epidemiologists usually do with age.

With this reframe, what additional studies are needed of comorbidity scores? Their comparative utility and convenience in exploratory analysis needs to be directly tested in identical populations predicting the same endpoints. Are they almost as good as expensive case-mix adjustment packages at guiding the investigator on the path to full rigorous control for confounding by multiple variables? How much residual confounding is apparent in score-adjusted RR, based on the change in RR upon using more rigorous methods to control confounding by comorbidities?

In conclusion, comorbidity scores (1) can simplify the data analysis process and might be useful for exploring confounding, (2) are unlikely to be good confounder adjustments despite their popularity, (3) do not standardize confounder adjustment across studies. Published evaluations of the predictive performance of comorbidity indices are limited and more cross-validation of different indices are needed to understand their utility.

Acknowledgements

Supported by grants from the Deutsche Forschungsgemeinschaft (DFG#Schn527/3-1 and DFG#Schn527/4-1) and the Pharmacoepidemiology Training and Research Grant, Harvard University.

References

- 1 Iezzoni LI. *Risk Adjustment for Measuring Healthcare Outcomes*. 2nd Edn. Chicago: Health Administration Press, 1997.
- 2 Goldfield N. *Physician Profiling and Risk Adjustment*. 2nd Edn. Gaithersburg: Aspen Publication, 1999.
- 3 Park RE, Brook RH, Koseoff J *et al*. Explaining variation in hospital death rates. *JAMA* 1990;**264**:484-90.
- 4 Greenfield S, Aronow HU, Elashoff RM, Watanabe D. Flaws in mortality data. *JAMA* 1988;**260**:2253-55.
- 5 Roos N, Wennberg JE, Malenka DJ *et al*. Mortality and reoperation after open and transurethral resection of the prostate for benign prostatic hyperplasia. *N Engl J Med* 1989;**320**:1120-24.
- 6 Malenka DJ, Roos N, Fisher ES *et al*. Further study of the increased mortality following transurethral prostatectomy; a chart based analysis. *J Urol* 1990;**144**:224-28.

- ⁷ Concato J, Horwitz RI, Feinstein AR, Elmore JG, Schiff SF. Problems of comorbidity in mortality after prostatectomy. *JAMA* 1992;**267**:1077-82.
- ⁸ Malcinka DJ, McLellan D, Roos N *et al.* Using administrative data to describe casemix: a comparison with the medical record. *J Clin Epidemiol* 1994;**47**:1027-32.
- ⁹ Hosmer DW, Lemeshow S. Confidence interval estimation on an index of quality performance based on logistic regression models. *Stat Med* 1995;**14**:2161-72.
- ¹⁰ Roos LL, Sharp SM, Cohen MM, Wajda A. Risk adjustment in claims-based research: the search for efficient approaches. *J Clin Epidemiol* 1989;**42**:1193-206.
- ¹¹ 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**:1075-79.
- ¹² 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.
- ¹³ Von Korff M, Wagner EH, Saunders K. A chronic disease score from automated pharmacy data. *J Clin Epidemiol* 1992;**45**:197-203.
- ¹⁴ 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.
- ¹⁵ D'Hoore W, Sicotte C, Tilquin C. Risk adjustment in outcome assessment: the Charlson Comorbidity Index. *Meth Inform Med* 1993;**32**:382-87.
- ¹⁶ 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.
- ¹⁷ Ghali WA, Hall RE, Rosen AK, Ash AS, Moskowitz MA. Searching for an improved clinical comorbidity index for use with ICD 9 CM administrative data. *J Clin Epidemiol* 1996;**49**:273-78.
- ¹⁸ Roos LL, Stranc L, James RC, Li Jianwei. Complications, comorbidities, and mortality: improving classification and prediction. *Health Serv Res* 1997;**32**:229-38.
- ¹⁹ Elixhauser A, Steiner C, Harris R, Coffey RM. Comorbidity measures for use with administrative data. *Med Care* 1998;**36**:8-27.
- ²⁰ Starfield B, Weiner J, Mumford L, Steinwachs D. Ambulatory care groups: a categorization of diagnoses for research and management. *Health Services Research* 1991;**26**:53-74.
- ²¹ 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.
- ²² Charlson ME, Szatrowski TP, Peterson J, Gold J. Validation of a combined comorbidity index. *J Clin Epidemiol* 1994;**47**:1245-51.
- ²³ Melfi C, Holleman E, Arthur D, Katz B. Selecting a patient characteristics index for the prediction of medical outcomes using administrative claims data. *J Clin Epidemiol* 1995;**48**:917-26.
- ²⁴ Poses RM, Smith WR, McClish DK, Anthony M. Controlling for confounding by indication for treatment. Are administrative data equivalent to clinical data? *Med Care* 1995;**33**:AS36-AS46.
- ²⁵ Clark DO, von Korff M, Saunders K, Baluch WM, Simon GE. A chronic disease score with empirically derived weights. *Med Care* 1995;**33**:783-95.
- ²⁶ Johnson RE, Hornbrook MC, Nichols GA. Replicating the chronic disease score (CDS) from automated pharmacy data. *J Clin Epidemiol* 1994;**47**:1191-99.
- ²⁷ Fowles JB, Lawthers AG, Weiner JP *et al.* Agreement between physician's office records and medicare part B claims data. *Health Care Fin Rev* 1995;**16**:189-99.
- ²⁸ Romano PS, Mark DH. Bias in the coding of hospital discharge data and its implications for quality assessment. *Med Care* 1994;**32**:81-90.
- ²⁹ Romano SR, Roos LL, Luft HS *et al.* A comparison of administrative versus clinical data: coronary artery bypass surgery as an example. *J Clin Epidemiol* 1994;**47**:249-60.
- ³⁰ Kieszak SM, Flanders WD, Kosinski AS, Shipp CC, Karp H. A comparison of the Charlson comorbidity index derived from medical records data and administrative billing data. *J Clin Epidemiol* 1999;**52**:137-42.
- ³¹ Redelmeier Donald A, Tan Siew H, Booth Gillian L. The treatment of unrelated disorders in patients with chronic medical diseases. *N Engl J Med* 1998;**338**:1516-20.
- ³² Glynn RJ, Monane M, Gurwitz JH, Chodnovskiy I, Avorn J. Agreement between drug treatment data and a discharge diagnosis of diabetes mellitus in the elderly. *Am J Epidemiol* 1999;**149**:541-49.
- ³³ 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.
- ³⁴ Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiol Rev* 1987;**9**:1-30.